TRANSCEIVER DESIGN FOR MIMO COMMUNICATIONS – A CHANNEL DECOMPOSITION PERSPECTIVE

By

YI JIANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2005

© Copyright by Yi Jiang, 2005

All rights reserved.

ACKNOWLEDGMENTS

Foremost, I thank my advisor, Professor Jian Li, for her support, encouragement, and guidance in the past four years. Dr. Li provided me the invaluable opportunity to investigate those fascinating research problems and always showed full confidence in me. I just hope I can live up to her expectations in my future career. I am very grateful to my collaborator, Professor William W. Hager, whose suggestions and rigorous math have benefited me a lot. I thank Dr. Tan F. Wong for teaching me the information theory. Some of the basic ideas of this dissertation were formulated when I was taking his class in the fall of 2003. I would like to thank Dr. John M. Shea, Dr. Tan F. Wong, Dr. Kenneth K. O, and Dr. William W. Hager for serving in my dissertation committee. Thanks go to all my friends both at the University of Florida and elsewhere who made the last four years full of fun.

This dissertation is dedicated to my parents and my fiancée Hongying.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

ABSTRACT ix

CHAPTER

1 INTRODUCTION 1

1.1 Two Categories of Schemes for MIMO Communications 1
1.2 Joint Transceiver Design: Where Tx and Rx Collaborate 2
1.3 MIMO Transceiver Design from Channel Decomposition Perspective . 3
1.4 Dissertation Outline 4

2 LINEAR MIMO TRANSCEIVER DESIGNS 6

2.1 Channel Model and Channel Capacity 6
2.1.1 Channel Model 6
2.1.2 Channel Capacity 8
2.2 Channel Capacity and Cramér-Rao Bound 10
2.3 Rate Performance of Linear Transceivers 13

3 MIMO TRANSCEIVER DESIGN USING GEOMETRIC MEAN DECOMPOSITION 17

3.1 VBLAST and ZF-DP 17
3.1.1 VBLAST 17
3.1.2 ZF-DP 18
3.2 Geometric Mean Decomposition for MIMO Transceiver Design 20
3.3 Performance Analyses and Implementations Issues 21
3.3.1 Performance Analyses 21
3.3.2 Combination of GMD with Two-way Channel Subspace Tracking 23
3.3.3 Subchannel Selection 24
3.3.4 Further Remarks 25
3.4 Performance Examples 27
3.5 Conclusions 28

4 UNIFORM CHANNEL DECOMPOSITION 33

4.1 Closed-Form Representation of MMSE-VBLAST 34
4.2 UCD-VBLAST 35
4.3 UCD-DP 39
4.4 Performance Analysis 40
4.4.1 Diversity Gain Analysis 40
4.4.2 Further Remarks 42

4.5 Numerical Examples 42
4.6 Conclusions 44

5 TUNABLE CHANNEL DECOMPOSITION 52

5.1 Introduction 52
5.2 Channel Model and Preliminaries 53
5.2.1 Channel Model 53
5.2.2 Channel Decomposition 53
5.2.3 Majorization and Generalized Triangular Decomposition 55
5.3 Tunable Channel Decomposition 57
5.3.1 TCD-VBLAST 57
5.3.2 TCD-DP 61
5.4 MIMO Communications with QoS Constraints 63
5.5 CDMA Sequence Design 70
5.5.1 CDMA Sequences Maximizing Sum Capacity 71
5.5.2 Uplink Case 71
5.5.3 Downlink Case 73
5.5.4 Numerical Example 73
5.5.5 Further Remarks 75
5.6 Conclusions 75

6 NOVEL MATRIX DECOMPOSITIONS 81

6.1 Introduction 81
6.2 Geometric Mean Decomposition 81
6.2.1 Generalized Maximin Properties 83
6.2.2 Implementation Based on Initial SVD 84
6.3 Generalized Triangular Decomposition 87
6.3.1 Existence of GTD 87
6.3.2 The GTD Algorithm 87
6.3.3 Inverse Eigenvalue Problem 93
6.4 Conclusions 95

7 CONCLUSIONS 99

REFERENCES 100

BIOGRAPHICAL SKETCH 105

LIST OF TABLES

Table page

4–1 The UCD-VBLAST scheme 38

5–1 The TCD-VBLAST Scheme 61

6–1 Comparison of SVD_EIG and GTD for inverse eigenvalue problems (CPU time in seconds, singular value and eigenvalue errors in sup-norm) . . . 93

LIST OF FIGURES

Figure page

3–1 Average capacity over 1000 Monte Carlo trials vs. SNR with $M_t = 4$ and $M_r = 4$ for i.i.d. Rayleigh flat fading channels. 29

3–2 Complementary cumulative distribution functions of the capacities of 5 subchannels of the i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 5$. Results based on 2000 Monte Carlo trials. 30

3–3 Complementary cumulative distribution function of the capacity of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$. Results based on 1000 Monte Carlo trials. SNR = (a) 0 dB, (b) 10 dB, (c) 20 dB, and (d) 30 dB. 31

3–4 BER performance averaged over 1000 Monte Carlo trials of i.i.d. Rayleigh flat fading channel vs. SNR with (a) $M_t = 2$ and $M_r = 4$ and (b) $M_t = 4$ and $M_r = 4$. 31

3–5 BER performances of GMD-VBLAST and GMD-ZFDP. Both are combined with OFDM for ISI suppression. 32

4–1 Complementary cumulative distribution function of the capacity of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$. Results based on 2000 Monte Carlo trials. SNR = (a) 10 dB, (b) 10 dB (c) 20 dB, and (d) 30 dB. 49

4–2 Complementary cumulative distribution functions of the capacities of 5 subchannels of an i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 5$. Results based on 2000 Monte Carlo trials. 50

4–3 Uncoded BER performance when using 16-QAM. Results based on 1000 Monte Carlo trials of an i.i.d. Rayleigh flat fading channel with $M_t = 4$ and $M_r = 4$. 50

4–4 BER performances of the UCD-DP, UCD-VBLAST schemes and the imaginary UCD-genie scheme. Results based on 1000 Monte Carlo trials of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$. 51

5–1 Illustration of the capacity lossless region obtainable via TCD. We assume $K = 3$, $C_1 = 3$, $C_2 = 2$, and $C_3 = 1$. 57

5–2 A Matlab function to solve (5.49). 68

5–3 Input SNR vs. Output SINR. The result is based on the average of 500 Monte Carlo trials of a i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 6$. 69

5–4 Input SNR vs. C_1. A rank 2 channel is decomposed into two subchannels with capacities C_1 and $C_2 = 10 - C_1$. 70

6–1 The operation displayed in (6.8) 86

6–2 The operation displayed in (6.15) 89

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

TRANSCEIVER DESIGN FOR MIMO COMMUNICATIONS
– A CHANNEL DECOMPOSITION PERSPECTIVE

By

Yi Jiang

May 2005

Chair: Jian Li
Major Department: Electrical and Computer Engineering

This dissertation studies the signal processing aspect of multi-input multi-output (MIMO) communications. The contribution of this dissertation is twofold.

First, this dissertation presents a new perspective to the MIMO communications: any MIMO scheme can be regarded as a MIMO channel decomposer, which decomposes (in an information lossy or lossless manner) a MIMO channel into multiple scalar subchannels. Based on this perspective, this dissertation presents three novel MIMO transceiver designs, the geometric mean decomposition (GMD) scheme, the uniform channel decomposition (UCD) scheme, and the tunable channel decomposition (TCD) scheme. All these schemes deploy either a decision feedback equalizer (DFE) at the receiver or a dirty paper precoder (DPP) at the transmitter. These transceiver designs represent a paradigm shift from the conventional linear MIMO transceiver designs to the nonlinear ones. The superior performance of the GMD and UCD schemes unveils the practical significance of making transmitter and receiver cooperate with each other. That is, such cooperations facilitate achieving the optimal tradeoff between the diversity gain and multiplexing promised by the MIMO communication theory. The TCD scheme represents a unifying solution to a considerably wide range of problems, including designing the precoder for orthogonal frequency division multiplexing (OFDM) communications and the optimal code division multiple access (CDMA) sequence design.

Second, this dissertation introduces two novel matrix decomposition algorithms, i.e., the geometric mean decomposition (GMD) and the generalized triangular decomposition (GTD). The two matrix decompositions form the cornerstones of the three transceiver

designs proposed in this dissertation. Moreover, the two decompositions have significant implications in the matrix analysis community. For instance, the GTD is a new solution to the inverse eigenvalue problem.

CHAPTER 1
INTRODUCTION

1.1 Two Categories of Schemes for MIMO Communications

Communications over multiple-input multiple-output (MIMO) wireless channels have been a subject of intense research over the past several years because deploying multiple antennas at both transmitter and receiver sides can drastically improve the spectral efficiency [1] [2] [3] [4]. For example, in contrast to the conventional additive white Gaussian noise (AWGN) channel whose spectral efficiency is

$$C(\mathsf{snr}) = \log_2(1 + \mathsf{snr}) \quad \text{bps/Hz},$$

without requiring additional input power, the MIMO channel with M_t transmitting antennas and M_r receiving antennas can have spectral efficiency as large as [1] [2]

$$C(\mathsf{snr}) = \min(M_r, M_t) \log_2(\mathsf{snr}) + O(1) \quad \text{bps/Hz},$$

given that there is plenty of scattering in the channel. Many spatial multiplexing methods, e.g., the BLAST scheme [2] [5] [6] [7] [8] [9] [10] [11], have been proposed to reap the great channel capacity.

Improving the data transmission reliability is another advantage of applying multiple antennas in wireless communications. By transmitting the same information through more than one independent fading channel, one can obtain much more reliable communications thanks to the redundance introduced. The space-time coding methods are based on such a rationale, (see, e.g., [12] [13] [14] [15]).

Zheng and Tse [16] show that one can exploit the diversity gain and multiplexing gain promised by the MIMO channel simultaneously. However, there is a fundamental tradeoff between the two gains. Zheng and Tse's theory provides a unifying framework to measure the performance of any MIMO schemes. Hence designing practical schemes capable of achieving the optimal diversity-multiplexing tradeoff is a central research topic in MIMO communications.

1.2 Joint Transceiver Design: Where Tx and Rx Collaborate

All the aforementioned methods assume that the channel state information (CSI) is available at the receiver (CSIR) only. Under this assumption, collaborations between the transmitter and receiver are difficult in the physical layer. However, if the communication environment is relatively stationary, the availability of CSI at the transmitter (CSIT) is also possible via feedback or the reciprocal principle when time division duplex (TDD) is used. In fact, in the third generation WCDMA standard [17], the CSIT is assumed to obtain improved system performance, which is referred to as the closed-loop transmit diversity or transmit adaptive array (TxAA) technique. Based on this assumption, the joint optimal transceiver design (also referred to as precoding at the transmitter and equalization at the receiver) has recently attracted considerable attentions [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29].

These designs are based on a variety of criteria, including minimum mean-squared-error (MMSE), [18] [21] [22], maximum SNR [21], maximum information rate [19] [20] [22], and BER based criteria [23] [24] [25] [29]. More recently, a unified framework has been presented to accommodate all these criteria, under which the design problems can be solved via convex optimization methods [26].

The aforementioned literature on joint transceiver design considered linear transformations only. It is widely understood that the singular value decomposition (SVD), which decomposes a MIMO channel into multiple parallel subchannels, and water filling can be used to achieve the channel capacity [3]. However, due to the usually very different signal-to-noise ratios (SNR) of the subchannels, this apparently simple scheme requires careful bit allocation (see, e.g., [19] [20] [23]) to match the subchannel capacity and achieve a prescribed BER. Bit allocation not only increases the coding/decoding complexity, but also is inherently capacity lossy because of the finite constellation granularity. An alternative is to use the same constellation in all the subchannels, like the schemes adopted by the European standard HIPERLAN/2 and the IEEE 802.11 standards for wireless local area networks (WLANs). However, for this alternative, the BER is dominated by the subchannels with the lowest SNRs. To optimize the BER performance, more signal power could be allocated to the poorer subchannels. Yet this approach causes significant capacity loss due to "inverse water filling" like power allocation. There is apparently a fundamental tradeoff between the capacity and the BER performance.

1.3 MIMO Transceiver Design from Channel Decomposition Perspective

In this dissertation, we present a new perspective to the MIMO communications. We regard the aforementioned MIMO schemes as MIMO channel decomposers, which decompose (in an information lossy or lossless manner) a MIMO channel into multiple scalar subchannels. For instance, the MIMO transceiver design based on SVD decomposes a MIMO into multiple eigen-subchannels. Similarly, the V-BLAST scheme decomposes a MIMO channel into multiple scalar subchannels which are referred to as *layers* by its inventors. These channel decompositions, however, are totally determined by the specific channel realization and one can have little control over how the channel is decomposed. For example, the gains of the subchannels obtained via SVD are totally determined by the singular values of the channel matrix, which one can have no control over.

An interesting question arises: if the transmitter and receiver are allowed to collaborate, how can we design a transceiver that can decompose a MIMO channel into multiple subchannels with prescribed channel gain, and without incurring capacity loss? This dissertation is devoted to answering this question. In the process of pursuing the answer, we investigate the following aspects of the problem.

First, we show that the conventional linear transceivers are inherently inflexible, and we cannot rely on linear transceivers to achieve our desired channel decompositions. Hence we need to go beyond the linearity constraint and investigate the nonlinear schemes, such as a decision feedback equalizer (DFE) and a dirty paper precoder (DPP).

Second, we study the possibility of new matrix decompositions other than using SVD. We propose two novel matrix decomposition algorithms, the geometric mean decomposition (GMD) and the generalized triangular decomposition (GTD). The two decompositions represent a wide class of matrix decomposition, which has significant implications in the matrix analysis community. For instance, the GTD is a new solution to the inverse eigenvalue problem.

Third, we propose three transceiver designs which combine the new matrix decomposition algorithms with the DFE and DPP. The three designs are the GMD scheme, the uniform channel decomposition (UCD) scheme and the tunable channel decomposition (TCD) scheme. Among them, the UCD scheme can decompose, in a strictly capacity lossless manner, a MIMO channel into multiple subchannels with *identical* capacities or, equivalently, identical channel gains. Moreover, the UCD scheme is a practical scheme

that can achieve the optimal tradeoff between the diversity gain and multiplexing gain. Without incurring any capacity loss, the TCD scheme can decompose a MIMO channel into multiple subchannels with prescribed capacities/channel gains. This scheme is applicable to a wide range of applications, including the multi-task communications where independent data streams with different qualities-of-service (QoS) share the same MIMO channel, and designing the optimal CDMA sequences.

1.4 Dissertation Outline

In Chapter 2, we introduce the data model and some relevant information-theoretic results that will be used in this dissertation. We also review the existing transceiver designs and analyze the performances of those methods. By linking the channel capacity with the Cramér Rao bound (CRB), we give an information-theoretic explanation why linear transceivers are inflexible.

Chapter 3 presents the GMD scheme that combines the VBLAST detector or DP precoder with the GMD matrix decomposition algorithm. The GMD scheme can decompose a MIMO channel into multiple *identical* scalar subchannels. This desirable property can bring much convenience to the practical system design, particularly the symbol constellation selection. Moreover, we have shown that the GMD scheme is optimal asymptotically for high SNR in terms of both information rate and BER performance while the computational complexity of our scheme is comparable to the conventional linear transceiver scheme.

In Chapter 4, we propose a uniform channel decomposition (UCD) scheme. Similar to the GMD scheme, the UCD is also based on the GMD matrix decomposition algorithm and can decompose a MIMO channel into multiple *identical* subchannels. Two remarkable merits of UCD, which are not shared by the GMD scheme, are that first, UCD is *strictly* capacity lossless at any SNR, and second, UCD can achieve the optimal diversity and multiplexing tradeoff. Moreover, the UCD scheme can decompose a MIMO channel into an arbitrarily large number of independent subchannels, which is an enabling technology to achieve high data rate transmission using small symbol constellations.

Chapter 5 is devoted to tackling a new aspect of the MIMO transceiver design problem. Instead of attempting to optimize the BER performance for fixed input power and data rate, we propose the TCD scheme which can decompose a MIMO channel into multiple subchannels with prescribed channel capacities. We show that TCD is a

solution to a wide range of applications, including the applications in which independent data streams with different qualities-of-service (QoS) share the same MIMO channel and design the optimal CDMA sequences.

The mathematical foundations of this dissertation, the GMD and GTD algorithms, are established in Chapter 6. The two novel matrix decomposition algorithms are the cornerstones of the MIMO transceiver designs proposed in this dissertation.

The conclusions are given in Chapter 7.

To read this dissertation, it is unnecessary to plunge into the details of the GMD and GTD algorithms. For this reason, we put them to the latter part of the dissertation. However, a rough understanding of the two algorithms is necessary to appreciate Chapters 3–5.

CHAPTER 2
LINEAR MIMO TRANSCEIVER DESIGNS

2.1 Channel Model and Channel Capacity

2.1.1 Channel Model

We consider a communication system with M_t transmitting and M_r receiving antennas in a frequency flat fading channel. The sampled baseband signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{F}\mathbf{x} + \mathbf{z}, \tag{2.1}$$

where $\mathbf{x} \in \mathbb{C}^{L\times 1}$ is the information symbols precoded by the precoder $\mathbf{F} \in \mathbb{C}^{M_t\times L}$ and $\mathbf{y} \in \mathbb{C}^{M_r\times 1}$ is the received signal and $\mathbf{H} \in \mathbb{C}^{M_r\times M_t}$ is the channel matrix with rank K. We assume $E[\mathbf{x}\mathbf{x}^*] = \sigma_x^2\mathbf{I}_L$ and $\mathbf{z} \sim N(0, \sigma_z^2\mathbf{I}_{M_r})$ is the circularly symmetric complex Gaussian noise, where $\mathbf{I}_L$ stands for an identity matrix with dimension L. We define the input SNR as

$$\rho = \frac{E[\mathbf{x}^*\mathbf{F}^*\mathbf{F}\mathbf{x}]}{\sigma_z^2} = \frac{\sigma_x^2}{\sigma_z^2}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\} \triangleq \frac{1}{\alpha}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\}, \tag{2.2}$$

where $\alpha = \dfrac{\sigma_x^2}{\sigma_z^2}$. Designing the MIMO transceivers, including the precoder $\mathbf{F}$ and the associated equalizer, is the focus of this dissertation.

We note that the data model in (2.1) is generic. For an intersymbol-interference (ISI) channel with impulse response $\mathbf{h} = [h_0, h_1, \ldots, h_{M-1}]^\mathsf{T}$ with $(\cdot)^\mathsf{T}$ denoting transpose, if a block data with length N are transmitted using the "zero-padded" OFDM, then the received block data can also be written in the form of (2.1) with

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & 0 & \ldots & 0 \\ \vdots & h_0 & 0 & \ldots & 0 \\ h_{M-1} & \ldots & \ddots & \ldots & \vdots \\ 0 & \ddots & \ldots & \ddots & 0 \\ \vdots & 0 & h_{M-1} & \ldots & h_0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & h_{M-1} \end{bmatrix}. \tag{2.3}$$

In this case, $\mathbf{H}$ is a Toeplitz matrix with its dimensionality $M_t = N$ and $M_r = N+M-1$. If the OFDM with cyclic prefix is used, the channel matrix is a circulant Toeplitz matrix,

i.e.,

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & h_{M-1} & \dots & \dots & h_1 \\ h_1 & h_0 & 0 & h_{M-1} & \ddots & h_2 \\ \vdots & \ddots & \ddots & 0 & \ddots & \vdots \\ h_{M-2} & \dots & \ddots & \ddots & \ddots & h_{M-1} \\ h_{M-1} & h_{M-2} & \dots & h_1 & h_0 & 0 \\ 0 & h_{M-1} & h_{M-2} & \dots & h_1 & h_0 \end{bmatrix}. \tag{2.4}$$

Here, $M_t = M_r = N$. In either case, if the block data are precoded with the linear precoder $\mathbf{F}$, then the received data are given in (2.1). This ISI channel problem has been studied in [21] [30].

In an idealized synchronous CDMA (S-CDMA) system where the channel does not experience any fading or near-far effect, L mobile users modulate their information symbols via spreading sequences $\{\mathbf{s}_i\}_{i=1}^{L}$, each of which has the processing gain N. The discrete-time baseband S-CDMA signal received at the (single-antenna) base-station can be represented as [31]

$$\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{z} \tag{2.5}$$

where $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_L] \in \mathbb{R}^{N \times L}$ and the lth ($1 \leq l \leq L$) entry of $\mathbf{x}$, x_l, stands for the information symbol from the lth user. In the downlink channel, the base station multiplexes the information dedicated to the L mobile users through the spreading sequences, which are the columns of $\mathbf{S}$. Then all the mobiles receive the same signal given in (2.5). We remark that (2.5) can also be written as (2.1) with $\mathbf{H} = \mathbf{I}_N$ and $\mathbf{F} = \mathbf{S}$. Here $M_r = M_t = N$ is the processing gain. Hence, optimizing the spreading sequences amounts to optimizing the precoder $\mathbf{F}$ for a MIMO system. Indeed, this problem has been under intensive research in the past several years.

In summary, both designing a precoder for OFDM transmission through an ISI channel and searching for the optimal S-CDMA sequences can be regarded as special cases in the unifying framework of MIMO transceiver designs. MIMO transceiver designs can be used in the OFDM and CDMA applications after only simple modifications. In this dissertation, we will concentrate on MIMO transceiver design although we will discuss the optimal design of CDMA sequences in Chapter 5.

2.1.2 Channel Capacity

Suppose $\mathbf{x}$ is a Gaussian random vector. The capacity of the MIMO channel (2.1) is [1]

$$C = \log_2 \frac{|\sigma_z^2 \mathbf{I} + \sigma_x^2 \mathbf{H}\mathbf{F}\mathbf{F}^*\mathbf{H}^*|}{|\sigma_z^2 \mathbf{I}|} \tag{2.6}$$

where $|\cdot|$ denotes the determinant of a matrix. If both CSIT and CSIR are available, we can maximize the channel capacity with respect to $\mathbf{F}$ given the input power constraint $\sigma_x^2 \mathrm{Tr}\{\mathbf{F}\mathbf{F}^*\} = \rho\sigma_z^2$. That is,

$$C_{IT} = \max_{\sigma_x^2 \mathrm{Tr}\{\mathbf{F}\mathbf{F}^*\} = \rho\sigma_z^2} \log_2 |\mathbf{I} + \alpha^{-1}\mathbf{H}\mathbf{F}\mathbf{F}^*\mathbf{H}^*|, \tag{2.7}$$

where α is as defined in (2.2) and the subscript of C_{IT} stands for "informed transmitter".

Denote the SVD of $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$, where $\mathbf{\Lambda}$ is a $K \times K$ diagonal matrix whose diagonal elements $\{\lambda_{H,k}\}_{k=1}^K$ are the nonzero singular values of $\mathbf{H}$. The solution to $\mathbf{F}$ in (2.7) is [3]

$$\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}. \tag{2.8}$$

Here $\mathbf{\Phi}$ is diagonal whose kth ($1 \le k \le K$) diagonal element ϕ_k determines the power loaded to the kth subchannel and is found via "water filling" to be

$$\phi_k(\mu) = \left(\mu - \frac{\alpha}{\lambda_{H,k}^2}\right)^+, \tag{2.9}$$

with μ being chosen such that $\sigma_x^2 \sum_{k=1}^K \phi_k(\mu) = \rho\sigma_z^2$ and $(a)^+ = \max\{0, a\}$. Then the solution to (2.7) is

$$C_{IT} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{\phi_k}{\alpha}\lambda_{H,k}^2\right) \text{ bps/Hz}. \tag{2.10}$$

Note that some of ϕ_k's can be zeros. In this case, we can only transmit $L < K$ data streams.

If the CSIT is not available, the optimal transmission strategy is to evenly allocate power to each antenna [3]. For this case, $\mathbf{F} = \sqrt{\frac{\rho\alpha}{M_t}}\mathbf{I}_{M_t}$ and the channel capacity with uninformed transmitter (UT) is

$$C_{UT} = \sum_{n=1}^{M_t} \log_2 \left(1 + \frac{\rho\lambda_{H,n}^2}{M_t}\right) \text{ bps/Hz}. \tag{2.11}$$

[1] Throughout this dissertation, we assume that the coherent time of the channel goes to infinity. Hence advanced coding is applicable to approach the Shannon capacity.

It is proven [32] that if $K = M_t$

$$\frac{C_{IT}}{C_{UT}} \to 1 \quad \text{as} \quad \rho \to \infty. \tag{2.12}$$

We claim a stronger relationship as follows. [2]

Lemma 2.1.1 *For the data model in (2.1), if the channel matrix* $\mathbf{H}$ *is of full column rank, i.e.,* $K = M_t$*, then*

$$C_{IT} - C_{UT} \to 0 \quad \text{as} \quad \rho \to \infty. \tag{2.13}$$

Proof: Inserting (2.9) into (2.10) yields

$$C_{IT} = \sum_{n=1}^{K} \log_2 \left(\mu \lambda_{H,n}^2\right)^+ \tag{2.14}$$

where μ is chosen such that

$$K\mu - \sum_{n=1}^{K} \frac{1}{\lambda_{H,n}^2} = \rho, \tag{2.15}$$

or

$$\mu = \frac{\rho}{K} + \frac{1}{K} \sum_{n=1}^{K} \frac{1}{\lambda_{H,n}^2}. \tag{2.16}$$

Here we assume that all the K subchannels are used because of large ρ, i.e., $\mu - \frac{1}{\lambda_{H,n}^2} > 0$ for $n = 1, 2, \ldots, K$.

From (2.14), (2.16), and (2.11), and using the fact that $K = M_t$, we have

$$C_{IT} - C_{UT} = \sum_{n=1}^{K} \log_2 \left(\frac{\rho + \sum_{n=1}^{K} \frac{1}{\lambda_{H,n}^2}}{\rho + \frac{K}{\lambda_{H,n}^2}} \right). \tag{2.17}$$

Note that

$$\lim_{\rho \to \infty} \frac{\rho + \sum_{n=1}^{K} \frac{1}{\lambda_{H,n}^2}}{\rho + \frac{K}{\lambda_{H,n}^2}} = 1 \quad \text{for} \quad 1 \le n \le K \tag{2.18}$$

and that $f(x) = \log_2 x$ is a continuous function if $x > 0$. The lemma follows immediately from (2.17). ∎

However, we note that CSIT can be very helpful in the following cases:

A. The SNR is low or moderate.

B. $\mathbf{H}$ is rank deficient or ill-conditioned.

[2] A similar, but somewhat vague, statement is found in [8].

C. There are more transmitting antennas than receiving ones, i.e., $M_t > M_r$.

Moreover, the availability of CSIT provides more freedom, which makes it easier to devise joint transceiver design schemes to achieve the underlying channel capacity. This observation is the underlying theme of this dissertation.

2.2 Channel Capacity and Cramér-Rao Bound

One of the most important significances of the Shannon's information theory is that this theory can predict the highest achievable data rate for a given channel. Similarly, the Cramér-Rao bound (CRB) [33], which is the inverse of the Fisher information matrix (FIM), can predict the minimum mean squared error (MMSE) an estimator can achieve. In this section, we show that the MIMO channel capacity formula of (2.6) can be re-written as a function of CRB, or FIM. Based on this relationship, we show that linear transceivers lack flexibility.

We rewrite (2.1) as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{2.19}$$

but we relax the assumption of (2.1) slightly. Instead of assuming spatially white noise, we assume that $\mathbf{z} \sim N(\mathbf{0}, \mathbf{R}_z)$. We also assume that the channel input $\mathbf{x} \sim N(\mathbf{0}, \mathbf{R}_x)$ also has circularly symmetric complex Gaussian distribution and is independent of $\mathbf{z}$. Then the channel output $\mathbf{y} \sim N(\mathbf{0}, \mathbf{H}\mathbf{R}_x\mathbf{H}^* + \mathbf{R}_z)$. For this more general scenario, the channel capacity is

$$C = \log \frac{|\mathbf{R}_z + \mathbf{H}\mathbf{R}_\mathbf{x}\mathbf{H}^*|}{|\mathbf{R}_z|} \tag{2.20}$$

Now Consider the following random vector,

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{R}_x & \mathbf{R}_x\mathbf{H}^* \\ \mathbf{H}\mathbf{R}_x & \mathbf{R}_y \end{bmatrix}\right) \tag{2.21}$$

Its log-likelihood function is

$$\log f(\mathbf{x}, \mathbf{y}) = -\text{const} - [\mathbf{x}^* \ \mathbf{y}^*] \begin{bmatrix} \mathbf{R}_x & \mathbf{R}_x\mathbf{H}^* \\ \mathbf{H}\mathbf{R}_x & \mathbf{R}_y \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{2.22}$$

Using the block matrix inversion formula [34], we get

$$\begin{bmatrix} \mathbf{R}_x & \mathbf{R}_x\mathbf{H}^* \\ \mathbf{H}\mathbf{R}_x & \mathbf{R}_y \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \diamond \end{bmatrix} \tag{2.23}$$

where

$$\mathbf{A} = \left(\mathbf{R}_x - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x\right)^{-1} \tag{2.24}$$

$$\mathbf{B} = -\left(\mathbf{R}_x - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x\right)^{-1}\mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1} \tag{2.25}$$

and $\diamond$ is irrelevant to the present discussion. From (2.22)-(2.25) we have

$$-\frac{\partial \log f(\mathbf{x},\mathbf{y})}{\partial \bar{\mathbf{x}}} = \left(\mathbf{R}_x - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x\right)^{-1}\left(\mathbf{x} - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{y}\right). \tag{2.26}$$

where $\bar{\mathbf{x}}$ is the conjugate of $\mathbf{x}$. Here we define the differentiation with respect to a complex-valued vector as [35, Appendix B]

$$\frac{\partial}{\partial \mathbf{w}} = \frac{1}{2}\begin{pmatrix} \frac{\partial}{\partial u_1} - j\frac{\partial}{\partial v_1} \\ \vdots \\ \frac{\partial}{\partial u_M} - j\frac{\partial}{\partial v_M} \end{pmatrix}, \qquad \frac{\partial}{\partial \bar{\mathbf{w}}} = \frac{1}{2}\begin{pmatrix} \frac{\partial}{\partial u_1} + j\frac{\partial}{\partial v_1} \\ \vdots \\ \frac{\partial}{\partial u_M} + j\frac{\partial}{\partial v_M} \end{pmatrix} \tag{2.27}$$

where the mth entry of $\mathbf{w}$, $w_m = u_m + jv_m$, $m = 1, \ldots, M$. The Bayesian Fisher information matrix (FIM) [36] is given by

$$\text{FIM} = E\left[\frac{\partial \log f(\mathbf{x},\mathbf{y})}{\partial \bar{\mathbf{x}}}\frac{\partial \log f(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}^{\mathsf{T}}\right]. \tag{2.28}$$

Based on (2.26) and (2.21), we obtain

$$\begin{aligned} \text{FIM} &= \mathbf{A}[\mathbf{I} \vdots -\mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}]\begin{bmatrix} \mathbf{R}_x & \mathbf{R}_x\mathbf{H}^* \\ \mathbf{H}\mathbf{R}_x & \mathbf{R}_y \end{bmatrix}\begin{bmatrix} \mathbf{I} \\ -\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x \end{bmatrix}\mathbf{A} \\ &= \left(\mathbf{R}_x - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x\right)^{-1} && (2.29) \\ &= \mathbf{R}_x^{-1} + \mathbf{H}^*(\mathbf{R}_y - \mathbf{H}\mathbf{R}_x\mathbf{H}^*)^{-1}\mathbf{H} && (2.30) \\ &= \mathbf{R}_x^{-1} + \mathbf{H}^*\mathbf{R}_z^{-1}\mathbf{H} && (2.31) \end{aligned}$$

Comparing (2.20) and (2.31), we see that

$$\begin{aligned} R &= \log_2|\mathbf{R}_x| + \log_2|\text{FIM}| && (2.32) \\ &= \log_2|\mathbf{R}_x| - \log_2|\text{CRB}| && (2.33) \end{aligned}$$

where

$$\text{CRB} = \text{FIM}^{-1} = \mathbf{R}_x - \mathbf{R}_x\mathbf{H}^*\mathbf{R}_y^{-1}\mathbf{H}\mathbf{R}_x. \tag{2.34}$$

This shows that there exists a simple relation between the Gaussian MIMO channel throughput, which is an upper bound of the information transmission rate for any

coder/decoder, and the CRB, which is a lower bound on the covariance matrix of any unbiased estimator of $\mathbf{x}$.

The MMSE estimator of $\mathbf{x}$ is

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathbf{R}_x\mathbf{H}\left(\mathbf{H}\mathbf{R}_x\mathbf{H}^* + \mathbf{R}_z\right)^{-1}\mathbf{y} \tag{2.35}$$

It is easy to verify that the MMSE estimator of $\mathbf{x}$ can achieve the CRB. Hence the MMSE estimator is the best we can achieve under the Gaussian assumptions. In general cases, the matrices FIM and CRB are non-diagonal; i.e., the MMSE estimates of the elements of $\mathbf{x}$ are correlated. The correlations between the elements of $\mathbf{x}$ clearly contain useful information for the subsequent decoding procedures. However, in practice, we only estimate the single elements of $\mathbf{x}$ separately and ignore the correlations between these elements. This causes the loss of information. In fact, we can quantify the capacity loss as

$$C_{\text{loss}} = \sum_{k=1}^{M_t} \log \text{CRB}_{kk} - \log|\text{CRB}| \tag{2.36}$$

where CRB_{kk} denotes the k-th diagonal element of CRB. According to the Hadamard inequality [34], for any positive semidefinite matrix $\mathbf{M} \in \mathbb{C}^K$,

$$|\mathbf{M}| \leq \prod_{i=1}^{K} M_{kk} \tag{2.37}$$

and the equality holds if and only if $\mathbf{M}$ is diagonal. Hence $C_{\text{loss}} \geq 0$ and there is no capacity loss if and only if CRB is a diagonal matrix.

Based on the aforementioned discussions, we see that i) in general MIMO communications, linear MMSE estimators followed by separate substream decoding are not capacity-wise optimal and ii) if the channel matrix $\mathbf{H}$ has the property that CRB of (2.34) is a diagonal matrices, linear MMSE estimators may be the first step of capacity lossless processing. If CSIT is available, the transmitter can apply some precoder $\mathbf{F}$ and get a virtual channel matrix

$$\mathbf{H}_{\text{vt}} = \mathbf{H}\mathbf{F} \tag{2.38}$$

such that CRB is diagonal. This explains why all the existing linear transceiver designs invariably lead to the diagonalization of the channel matrix. Indeed, if $\mathbf{R}_x$ is diagonal and $\mathbf{R}_z = \sigma_z^2\mathbf{I}$, then it follows from (2.31) that $\mathbf{H}_{\text{vt}}$ must have orthogonal columns to get diagonal FIM and hence diagonal CRB. Then the precoder $\mathbf{F} = \mathbf{V}$, which is the right singular vector of $\mathbf{H}$, is the only optimal solution. Yet as we discussed

before, this inflexible transceiver scheme can bring many difficulties to the subsequent coding/decoding and modulation/demodulation procedures.

2.3 Rate Performance of Linear Transceivers

To gain more insights into the limitations of the linear transceiver designs, we analyze the asymptotic rate performances of two typical linear transceiver designs for high SNR. We will show that the linear transceivers may suffer from considerable capacity loss and there is apparently a fundamental tradeoff between the throughput and the BER performance.

According to the channel model of (2.1), the received data vector is

$$\mathbf{y} = \mathbf{HFx} + \mathbf{z}. \tag{2.39}$$

The optimal linear receiver is always the LMMSE equalizer (also see, e.g., [23])

$$\mathbf{G}_{opt} = \mathbf{F}^*\mathbf{H}^* \left(\mathbf{HFF}^*\mathbf{H}^*\sigma_x^2 + \sigma_z^2\mathbf{I}\right)^{-1}, \tag{2.40}$$

which yields the optimal estimate of the information symbol $\hat{\mathbf{s}} = \mathbf{G}_{opt}\mathbf{y}$. The mean-squared-error (MSE) matrix of $\hat{\mathbf{s}}$ is

$$\mathbf{E} = \left(\mathbf{I} + \alpha^{-1}\mathbf{F}^*\mathbf{H}^*\mathbf{HF}\right)^{-1}. \tag{2.41}$$

Note that $\mathbf{E}$ is a function of the linear precoder $\mathbf{F}$. In the following, we analyze two linear precoder designs based on the minimization of the trace of the MSE matrix (MTM) and the minimization of the maximum diagonal elements of MSE matrix (MMD) criteria, which are referred to as ARITH-MSE and MAX-MSE in [26], respectively. We choose these two schemes because they appear to be the most typical ones and the MMD scheme yields the optimal (or very close to the optimal) performance among all the linear transceivers. Indeed, the MMD is equivalent to the linear MIN-BER scheme in the flat fading channel case (see [26]). We do not consider the SVD plus water filling scheme herein since it requires the complicated bit loading.

The MTM scheme, or ARITH-MSE, which has appeared in several linear transceiver design papers (see, e.g., [22] [23] and [26]), attempts to minimize tr($\mathbf{E}$) with respect to $\mathbf{F}$. The MTM precoder turns out to be

$$\mathbf{F}_{MTM} = \mathbf{V}\mathbf{\Phi}^{1/2} \tag{2.42}$$

where $\mathbf{V}$ is as defined in the SVD $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$, and $\mathbf{\Phi}$ is a diagonal matrix whose ith diagonal element ϕ_i denotes the signal power loaded to the ith subchannel. According to the literature (see e.g. [23] Sec. III-A)

$$\phi_i = \left(\frac{\sigma_z}{\mu^{1/2}\lambda_{H,i}} - \frac{\sigma_z^2}{\lambda_{H,i}^2} \right)^+, \tag{2.43}$$

where μ is the Lagrange multiplier which controls the loaded power such that $\sum_{i=1}^{K} \phi_i = \rho\sigma_z^2$. Suppose ρ is sufficiently large. Then all the K subchannels are used and

$$\sum_{i=1}^{K} \left(\frac{\sigma_z}{\mu^{1/2}\lambda_{H,i}} - \frac{\sigma_z^2}{\lambda_{H,i}^2} \right) = \rho\sigma_z^2 \tag{2.44}$$

or

$$\mu^{-1/2} = \frac{\rho + \sum_{i=1}^{K} \lambda_{H,i}^{-2}}{\sum_{i=1}^{K} \lambda_{H,i}^{-1}\sigma_z^{-1}}. \tag{2.45}$$

Substituting (2.42), (2.43) and (2.45) into (2.41), we see that $\mathbf{E}$ is diagonal with the ith diagonal element

$$E_i = \frac{\sum_{k=1}^{K} \lambda_{H,k}^{-1}}{(\rho + \sum_{k=1}^{K} \lambda_{H,k}^{-2})\lambda_{H,i}}, \tag{2.46}$$

Then (cf. Equation (28) of [26])

$$\begin{aligned} C_i &= -\log_2 E_i & (2.47)\\ &= \log_2 \left(\frac{\rho + \sum_{k=1}^{K} \lambda_{H,k}^{-2}}{\sum_{i=k}^{K} \lambda_{H,k}^{-1}} \right) + \log_2 \lambda_{H,i}. & (2.48) \end{aligned}$$

Hence the sum rate of the channel using the MTM scheme is

$$C_{MTM} = \sum_{i=1}^{K} C_i = K \log_2 \left(\frac{\rho + \sum_{k=1}^{K} \lambda_{H,k}^{-2}}{\sum_{k=1}^{K} \lambda_{H,k}^{-1}} \right) + \sum_{i=1}^{K} \log_2 \lambda_{H,i}. \tag{2.49}$$

The channel capacity with uniform power loading in the K subchannels is

$$C_{UPL} = \sum_{i=1}^{K} \log_2 \left(1 + \frac{\rho}{K} \lambda_{H,i}^2 \right). \tag{2.50}$$

Here C_{UPL} is different from C_{UT} defined in (2.11) in that C_{UPL} corresponds to the channel with the transmitter knowing the range space of $\mathbf{H}$.

It follows from (2.49) and (2.50) that

$$C_{UPL} - C_{MTM} = \sum_{i=1}^{K} \log_2 \left(1 + \frac{\rho}{K} \lambda_{H,i}^2 \right) - K \log_2 \left(\frac{\rho + \sum_{i=1}^{K} \lambda_{H,i}^{-2}}{\sum_{i=1}^{K} \lambda_{H,i}^{-1}} \right) - \sum_{i=1}^{K} \log_2 \lambda_{H,i}. \tag{2.51}$$

After some straightforward calculations, we have

$$\lim_{\rho\to\infty} C_{UPL} - C_{MTM} = K \log_2 \frac{\frac{1}{K}\sum_{i=1}^{K} \lambda_{H,i}^{-1}}{\left(\prod_{i=1}^{K} \lambda_{H,i}^{-1}\right)^{1/K}} \quad \text{bps/Hz}. \tag{2.52}$$

Note that for any real valued sequence $\{\lambda_{H,i}\}_{i=1}^{K} > 0$, the arithmetic mean is greater than or equal to the geometric mean, or $\frac{1}{K}\sum_{i=1}^{K} \lambda_{H,i}^{-1} \geq \left(\prod_{i=1}^{K} \lambda_{H,i}^{-1}\right)^{1/K}$. Hence we conclude that $\lim_{\rho\to\infty} C_{UPL} - C_{MTM} \geq 0$ and the equality holds if and only if $\{\lambda_{H,i}\}_{i=1}^{K}$ are all the same. We infer from (2.52) that the capacity loss of the MTM transceiver can be quite large if the channel matrix $\mathbf{H}$ has a large condition number, which is verified in Section 3.4.

If the same constellation is used for each subchannel, then the substream corresponding to the largest E_i dominates the overall BER performance. Recall that $E_i = \frac{\sum_{k=1}^{K} \lambda_{H,k}^{-1}}{(\rho+\sum_{k=1}^{K} \lambda_{H,k}^{-1})\lambda_{H,i}}$, which is proportional to the inverse of $\lambda_{H,i}$. Hence the subchannels may have very different SNRs especially when $\mathbf{H}$ has a large condition number. To mitigate this undesirable effect, one can use the MMD transceiver, or MAX-MSE (cf. [26] Section V-A5), with

$$\mathbf{F}_{MMD} = \mathbf{F}_{MTM}\mathbf{\Theta}, \tag{2.53}$$

where $\mathbf{\Theta}$ is a unitary matrix that makes all the diagonal elements of $\mathbf{E}$ in (2.41) the same, that is,

$$\bar{E} = \frac{1}{K}\sum_{i=1}^{K} E_i. \tag{2.54}$$

According to (2.47), the capacity of the channel using the MMD linear transceiver is

$$C_{MMD} = -K \log_2 \bar{E} = -K \log_2 \frac{1}{K}\sum_{i=1}^{K} E_i. \tag{2.55}$$

Thus

$$C_{MTM} - C_{MMD} = K \log_2 \frac{1}{K}\sum_{i=1}^{K} E_i - \sum_{i=1}^{K} \log_2 E_i \tag{2.56}$$

$$= K \log_2 \frac{\frac{1}{K}\sum_{i=1}^{K} E_i}{\left(\prod_{i=1}^{K} E_i\right)^{1/K}} \tag{2.57}$$

$$= K \log_2 \frac{\frac{1}{K}\sum_{i=1}^{K} \lambda_{H,i}^{-1}}{\left(\prod_{i=1}^{K} \lambda_{H,i}^{-1}\right)^{1/K}} \tag{2.58}$$

where to get (2.58) from (2.57) we have used (2.46). Note that the relative capacity loss of MMD compared with MTM is independent of SNR given that all the subchannels are used. Interestingly, we can see from (2.58) and (2.52) that $C_{MTM} - C_{MMD} = \lim_{\rho\to\infty} C_{UPL} - C_{MTM}$. We conclude that asymptotically for high SNR, the MMD transceiver has twice the capacity loss of MTM, i.e.,

$$\lim_{\rho\to\infty} C_{UPL} - C_{MMD} = 2K \log_2 \frac{\frac{1}{K}\sum_{i=1} \lambda_{H,i}^{-1}}{\left(\prod_{i=1}^{K} \lambda_{H,i}^{-1}\right)^{1/K}} \quad \text{bps/Hz}, \tag{2.59}$$

although it may yield better BER performance. An intuitive explanation of the capacity loss of the MMD transceiver is as follows. Note that the only difference between MTM and MMD is the prerotation matrix $\mathbf{\Theta}$, which is an invariant operator in terms of information capacity. However, $\mathbf{\Theta}$ makes the MSE matrix $\mathbf{E}$ non-diagonal, which means that the elements of $\hat{\mathbf{s}} = \mathbf{G}_{opt}\mathbf{y}$ are correlated. Clearly, the correlation contains useful information for symbol detection and decoding. However, the linear equalizer ignores the correlation, which results in the additional capacity loss quantified in (2.58). The analyses here are verified in Section 3.4.

In summary, the MTM transceiver suffers from capacity loss of (2.52) due to the information theoretically non-optimal power loading defined in (2.43). The MMD transceiver suffers from additional capacity loss because it makes the MSE matrix non-diagonal. Hence there is an apparently inevitable tradeoff between the information rate and BER performance if the same symbol constellation is used in the different subchannels. In the next chapter, we will introduce the GMD scheme and clarify that there is not necessarily a tradeoff between BER performance and channel capacity. Indeed, the GMD scheme attempts to achieve the best of both worlds simultaneously.

CHAPTER 3
MIMO TRANSCEIVER DESIGN USING GEOMETRIC MEAN DECOMPOSITION

3.1 VBLAST and ZF-DP

In this section, we first give a brief introduction to the VBLAST architecture [5], which is equivalent to the generalized decision feedback equalizer (GDFE) [37]. We also introduce the more recent zero-forcing "dirty paper" precoder (ZFDP) applied to the MIMO broadcast channels [38] [39].

3.1.1 VBLAST

VBLAST is a simple suboptimal receiver interface which is used in the MIMO system assuming that only CSIR is available. For a MIMO system (2.1) with $M_t \leq M_r$ and rank $K = M_t$, the transmitter allocates independent bit streams across the M_t transmitting antennas with no precoding. To decode the transmitted information symbol, VBLAST first estimates the signal with the spatial structure $\mathbf{h}_{M_t}$, where $\mathbf{h}_i$ denotes the ith column of $\mathbf{H}$, and then cancels it out from the received signal vector. Next, it estimates the signal with spatial structure $\mathbf{h}_{M_t-1}$ and so on. The signal estimator can be either the ZF or MMSE estimator. Some proper reordering of the columns of $\mathbf{H}$ is helpful to improve the BER performance [5]. This decoding scheme involves *sequential nulling and cancellation* which is proved to be equivalent to the generalized decision feedback equalizer (GDFE) [37].

The ZF nulling step in the VBLAST scheme can be represented by the QR decomposition $\mathbf{H} = \mathbf{QR}$ where $\mathbf{Q}$ is an $M_t \times K$ matrix with orthonormal columns and $\mathbf{R}$ is a $K \times K$ upper triangular matrix. Let us rewrite (2.1) as

$$\mathbf{y} = \mathbf{QRx} + \mathbf{z}. \tag{3.1}$$

Multiplying $\mathbf{Q}^*$ to both sides of (3.1) yields

$$\bar{\mathbf{y}} = \mathbf{Rx} + \bar{\mathbf{z}}, \tag{3.2}$$

or

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_K \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ 0 & r_{22} & \cdots & r_{2K} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{KK} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} + \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \vdots \\ \tilde{z}_K \end{bmatrix}. \tag{3.3}$$

The sequential signal detection is as follows

for $i = K : -1 : 1$

$\hat{x}_i = \mathcal{C}\left[\left(\tilde{y}_i - \sum_{j=i+1}^{K} r_{ij}\hat{x}_j\right)/r_{ii}\right]$

end

where $\mathcal{C}$ stands for mapping to the nearest symbol in the symbol constellation. Ignoring the error-propagation effect, we see that the MIMO channel is decomposed into K parallel scalar subchannels

$$\tilde{y}_i = r_{ii}x_i + \tilde{z}_i, \quad i = 1, 2, \ldots, K. \tag{3.4}$$

3.1.2 ZF-DP

We consider a broadcast MIMO channel with M_t transmitting antennas and M_r receiving antennas ($M_t \geq M_r$). The channel model is exactly the same as (2.1) and the CSIT is available. However the receiving antennas cannot cooperate with each other. A vector transmission scheme was proposed in [40], which combines the QR decomposition and "dirty paper" precoding. We refer to this approach as the zero-forcing "dirty paper" precoding (ZFDP). (The use of the "dirty paper" phrase is due to Costa [41].)

The ZFDP scheme resembles the zero-forcing VBLAST method. It also goes through the *sequential nulling and cancellation* procedure. The only difference is that all these operations are done by the transmitter.

By assuming $\mathbf{H}$ to be of full row rank, i.e., $K = M_r$, ZFDP also begins with the QR decomposition $\mathbf{H}^* = \breve{\mathbf{Q}}\breve{\mathbf{R}}$. Let us rewrite (2.1) as

$$\mathbf{y} = \breve{\mathbf{R}}^*\breve{\mathbf{Q}}^*\mathbf{x} + \mathbf{z}. \tag{3.5}$$

Denoting $\mathbf{x} = \breve{\mathbf{Q}}\breve{\mathbf{x}}$ yields

$$\mathbf{y} = \breve{\mathbf{R}}^*\breve{\mathbf{x}} + \mathbf{z}, \tag{3.6}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} \breve{r}_{11} & 0 & \dots & 0 \\ \breve{r}_{21} & \breve{r}_{22} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \breve{r}_{K1} & \dots & \dots & \breve{r}_{KK} \end{bmatrix} \begin{bmatrix} \breve{x}_1 \\ \breve{x}_2 \\ \vdots \\ \breve{x}_K \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} \tag{3.7}$$

Denote $\mathbf{s} \in \mathbb{C}^{K\times 1}$ to be the symbol vector destined for the K receivers. We wish to have $\breve{\mathbf{x}}$ satisfying

$$\begin{bmatrix} \breve{r}_{11}s_1 \\ \breve{r}_{22}s_2 \\ \vdots \\ \breve{r}_{KK}s_r \end{bmatrix} = \begin{bmatrix} \breve{r}_{11} & 0 & \dots & 0 \\ \breve{r}_{21} & \breve{r}_{22} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \breve{r}_{K1} & \dots & \dots & \breve{r}_{KK} \end{bmatrix} \begin{bmatrix} \breve{x}_1 \\ \breve{x}_2 \\ \vdots \\ \breve{x}_K \end{bmatrix} \tag{3.8}$$

The solution to (3.8) is

$$\breve{\mathbf{x}} = \breve{\mathbf{R}}^{-*}\text{diag}\{\breve{\mathbf{R}}\}\mathbf{s}. \tag{3.9}$$

However, the matrix inversion can amplify the norm of $\breve{\mathbf{x}}$ significantly which can lead to additional power consumption at the transmitter. By exploiting the finite alphabet property of the communication signals, the modulo arithmetic precoder (more recently known as the Tomlinson-Harashima Precoder [42], [43]) can be applied to bound the value of the transmitted signal. Moreover, the trellis precoding can be used to eliminate the 1.53 dB shape-loss of Tomlinson-Harashima precoding [44]. The ZFDP transmission scheme decomposes the MIMO channel into K parallel scalar channels (see [40] for more details)

$$y_i = \breve{r}_{ii}\breve{x}_i + z_i \qquad i = 1, 2, \dots, K. \tag{3.10}$$

Several remarks are now in order. a) VBLAST is shown to be able to achieve only about 72% of the capacity [5]. That is because imposing the same rate of transmission on all the transmitters makes the channel capacity limited by the worst of the K scalar subchannels. b) VBLAST has only diversity gain of $M_r - M_t + 1$. c) ZFDP can achieve the broadcast channel capacity for high SNR [39], but the subchannels have different fading levels. Hence the transmitter, just like the aforementioned linear transceivers, have to consider the tradeoffs between the BER performance and the channel throughput. d) ZFDP scheme causes no error propagation, and thus (3.10) is precise. e) Both VBLAST and ZFDP involve nonlinear operations.

3.2 Geometric Mean Decomposition for MIMO Transceiver Design

Note that VBLAST assumes no cooperations among transmitting antennas and ZFDP assumes no cooperations at the receivers. Then a natural question arises: can we exploit both the CSIR and CSIT to make things better if both CSIR and CSIT are available? We attempt to address this question next.

In the sequel, we assume that the same signal constellation is used in all the independent symbol streams to reduce the system complexity. This is consistent with the HIPERLAN/2 and IEEE 802.11 standards. Then the overall BER performance of the system will be limited by the subchannel with the lowest SNR. To mitigate this problem, based on (3.4) and (3.10), we consider the following optimization problem

$$\begin{aligned} \max_{\mathbf{Q},\mathbf{P}} \quad & \min \quad \{r_{ii} : 1 \le i \le K\} \\ \text{subject to} \quad & \mathbf{R} = \mathbf{Q}^*\mathbf{H}\mathbf{P} \\ & \mathbf{R} \in \mathbb{R}^{K\times K}, r_{ij} = 0 \ \text{ for } \ i > j \\ & r_{ii} > 0 \ \text{ for } \ 1 \le i \le K \\ & \mathbf{Q}^*\mathbf{Q} = \mathbf{P}^*\mathbf{P} = \mathbf{I}_K \end{aligned} \tag{3.11}$$

where the semi-unitary matrices $\mathbf{Q}$ and $\mathbf{P}$ denote the linear operations at the receiver and transmitter, respectively.

Since both $\mathbf{Q}$ and $\mathbf{P}$ are semi-unitary matrices, we have $\prod_{n=1}^{K} r_{nn} = \prod_{n=1}^{K} \lambda_{H,n}$, where $\{\lambda_{H,n}\}_{n=1}^{K}$ are the K non-zero singular values of $\mathbf{H}$. In Chapter 6 we show that if there exist semi-unitary matrices $\mathbf{P}$ and $\mathbf{Q}$ satisfying

$$\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^*, \qquad \text{or equivalently,} \qquad \mathbf{R} = \mathbf{Q}^*\mathbf{H}\mathbf{P} \tag{3.12a}$$

where the diagonal elements of $\mathbf{R}$ are given by

$$r_{ii} = \bar{\lambda}_H \triangleq \left(\prod_{n=1}^{K} \lambda_{H,n}\right)^{1/K}, \quad 1 \le i \le K, \tag{3.12b}$$

then the $\mathbf{R}$ in (3.12) is the solution to (3.11). The detailed treatment of the decomposition (3.12) is delegated to Chapter 6, We refer to this decomposition as the geometric mean decomposition (GMD) since the diagonal elements of $\mathbf{R}$ are the geometric mean of $\{\lambda_{H,n}\}_{n=1}^{K}$. A computationally efficient and numerically stable algorithm is proposed in Section 6.2 to calculate the decomposition.

It seems reasonable to constrain the linear equalizer $\mathbf{Q}$ to be semi-unitary since it will keep the background noise white. Yet it seems unnecessary to constrain $\mathbf{P}$ to be

semi-unitary as well. Indeed, the constraint that $\mathbf{P}$ and $\mathbf{Q}$ should be semi-unitary is in fact *inactive* as shown in the following lemma established in Section 6.2.1.

Lemma 3.2.1 *The GMD of (3.12) is also the solution to the following optimization problem with relaxed constraints:*

$$\begin{aligned} \max_{\mathbf{P},\mathbf{Q}} \ & \min \ \{r_{ii} : 1 \le i \le K\} \\ \textit{subject to} \quad & \mathbf{R} = \mathbf{Q}^*\mathbf{H}\mathbf{P}, \ \ r_{ij} = 0 \ \ \textit{for} \ \ i > j, \ \mathbf{R} \in \mathbb{R}^{K\times K}, \\ & r_{ii} > 0, \ \ 1 \le i \le K, \\ & \mathrm{tr}(\mathbf{Q}^*\mathbf{Q}) \le K, \ \ \mathrm{tr}(\mathbf{P}^*\mathbf{P}) \le K. \end{aligned} \tag{3.13}$$

Proof: Omitted. See Section 6.2.1 for details. ■

The GMD, which can be viewed as an extended QR decomposition, can be readily combined with the aforementioned VBLAST (GDFE) or ZFDP. GMD-VBLAST is implemented as follows: We first calculate the GMD $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^*$. Next we choose the precoder $\mathbf{F} = \mathbf{P}$, then the equivalent data model is

$$\mathbf{y} = \mathbf{Q}\mathbf{R}\mathbf{x} + \mathbf{z}. \tag{3.14}$$

The next step is nothing but the VBLAST detector.

Ignoring the error propagation effect, we can regard the resulting subchannels as K independent and *identical* subchannels

$$y_i = \bar{\lambda}_H x_i + z_i, \quad \text{for} \quad i = 1, \ldots, K. \tag{3.15}$$

The GMD-ZFDP scheme is similar to GMD-VBLAST because of the duality between VBLAST and ZFDP.

3.3 Performance Analyses and Implementations Issues

In this section, we first present the performance analyses of the GMD scheme from capacity perspective, from which we demonstrate the advantages of our GMD scheme over the linear transceivers. Next, we consider combining the GMD scheme with the blind two-way channel subspace tracking in the TDD scenario. To achieve close to optimal performance at low SNR, we propose to combine GMD with subchannel selection. Finally, we discuss the relationship between our GMD scheme and [30].

3.3.1 Performance Analyses

As we have mentioned earlier, the overall BER performance of a MIMO communication system is dominated by the worst subchannels asymptotically for high SNR.

Hence the scheme optimizing the worst subchannel can enjoy the optimal BER performance for high SNR. This observation is also the motivation of the aforementioned MMD scheme. As a major advantage over the linear transceiver schemes, the GMD scheme is also asymptotically optimal in terms of the channel capacity for high SNR as we will show below.

If the signal power is allocated evenly to the K subchannels, then based on (3.15), we get

$$C_{GMD} = K \log_2 \left(1 + \frac{\rho}{K} \bar{\lambda}_H^2\right), \tag{3.16}$$

where ρ is defined in (2.2). The channel capacity with uniform power loading on the K subchannels is (see (2.50))

$$C_{UPL} = \sum_{n=1}^{K} \log_2 \left(1 + \frac{\rho}{K} \lambda_{H,n}^2\right). \tag{3.17}$$

It follows from (3.16) and (3.17) that

$$C_{UPL} - C_{GMD} = \log_2 \frac{\prod_{n=1}^{K} \left(1 + \rho \lambda_{H,n}^2\right)}{\left(1 + \rho \bar{\lambda}_H^2\right)^K}. \tag{3.18}$$

From (3.12b) and (3.18), we have

$$\lim_{\rho \to \infty} C_{UPL} - C_{GMD} = 0. \tag{3.19}$$

Based on Lemma 2.1.1

$$\lim_{\rho \to \infty} C_{IT} - C_{UPL} = 0. \tag{3.20}$$

Hence, it follows from (3.19) and (3.20) that

$$\lim_{\rho \to \infty} C_{IT} - C_{GMD} = 0, \tag{3.21}$$

i.e., for high SNR the GMD scheme is asymptotically optimal.

Hence the GMD scheme does not need to make the tradeoff between the information rate and BER performance as the conventional linear transceivers. Instead, our GMD scheme can achieve the optimum on both aspects simultaneously for high SNR.

As we have mentioned before, VBLAST may suffer from error propagation. Hence the BER performance of GMD-VBLAST will be inferior to the scalar equivalence in (3.15). We calculate the upper bound of the GMD-VBLAST BER as follows. For a fixed SNR ρ, we assume that the system of (3.15) has symbol error rate (SER) P_e, i.e., each subchannel has SER P_e/K. We consider the worst case that decoding errors in some

subchannels will cause the failure of the decoding in all the subsequent subchannels. The SER upper bound is readily calculated as

$$\begin{aligned} P_{e,GMD-VBLAST} &= \frac{1}{K}\sum_{n=0}^{K-1}(1-P_e)^n(K-n)P_e \\ &< \frac{1}{K}\sum_{n=0}^{K-1}(K-n)P_e \\ &= \frac{K+1}{2}P_e. \end{aligned} \tag{3.22}$$

For a moderate K, say $K \leq 10$, the performance loss caused by the error propagation is rather small. For a system with high dimensionality, GMD-ZFDP is a better choice since it causes no error propagation. On the other hand, the Tomlinson-Harashima precoder leads to an input power increase of $\frac{M}{M-1}$ for M-QAM.

3.3.2 Combination of GMD with Two-way Channel Subspace Tracking

In TDD systems, the GMD scheme may be combined with two-way channel subspace tracking techniques. The GMD algorithm, given in Chapter 6, starts with the SVD. To calculate the matrix $\mathbf{P}$ (cf. (3.11)), we only need to know the singular values $\mathbf{\Lambda}$ and the right singular vectors $\mathbf{V}$ (cf. Chapter 6). Similarly, only $\mathbf{\Lambda}$ and $\mathbf{U}$ are used to calculate $\mathbf{Q}$. Rewriting (2.1) with the precoder $\mathbf{F} = \mathbf{P}$ yields,

$$\mathbf{y} = \mathbf{HPx} + \mathbf{z}. \tag{3.23}$$

Since the GMD scheme uses the same signal constellation and uniform power allocation, the covariance matrix of $\mathbf{s}$ is a scaled identity matrix, i.e., $E[\mathbf{xx}^*] = \sigma_x^2\mathbf{I}$. Hence,

$$\mathbf{R_y} = E[\mathbf{yy}^*] = \mathbf{HH}^*\sigma_x^2 + \sigma_z^2\mathbf{I}. \tag{3.24}$$

If the signal power σ_x^2 and the noise power σ_z^2 are known *a priori*, we have $\mathbf{HH}^* = (\mathbf{R_y} - \sigma_z^2\mathbf{I})/\sigma_x^2$. Applying SVD to $\mathbf{HH}^*$, we get

$$\mathbf{HH}^* = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^*. \tag{3.25}$$

The GMD algorithm can be applied based on $\mathbf{U}$ and $\mathbf{\Lambda}$ to get the matrices $\mathbf{Q}$ and $\mathbf{R}$, which are sufficient for decoding. If a TDD system is used, the reverse channel, where the roles of previous transmitter and receiver are exchanged, can be modeled as

$$\mathbf{y}_{rev} = \mathbf{H}^\mathrm{T}\mathbf{Q}^*\mathbf{s}_{rev} + \mathbf{z}_{rev} \tag{3.26}$$

where the subscript "rev" means "reverse channel". Define

$$\mathbf{R}_{\mathbf{y}_{rev}} = E[\bar{\mathbf{y}}_{rev}\mathbf{y}_{rev}^{\mathrm{T}}] \tag{3.27}$$

where $\bar{\mathbf{y}}$ denotes the complex conjugate of $\mathbf{y}$. Using the similar argument, we have

$$\mathbf{H}^*\mathbf{H} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^*. \tag{3.28}$$

Then the *reverse* receiver, i.e., the previous transmitter, can calculate $\mathbf{R}$ and $\mathbf{P}$ from $\mathbf{V}$ and $\mathbf{\Lambda}$. Channel subspace tracking techniques (see, e.g., [45] [46]) can be used to estimate $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Lambda}$ efficiently. Hence our GMD scheme can be applied without the need of using training symbols for channel estimation. We note that this merit of GMD is not shared by the conventional transceiver schemes introduced in Section 2.3 since all those methods allocate different powers to different subchannels, which makes it difficult, if not impossible, to estimate the singular values in $\mathbf{\Lambda}$. Of course, if the same power is allocated to each eigen-subchannel, this blind two-way channel subspace tracking idea can also be combined with the SVD based schemes, at the cost of significant capacity loss.

The GMD scheme can be made backward compatible with the TDD systems using VBLAST decoders. By using CSIT or blind subspace tracking techniques, the transmitter can calculate the linear precoder $\mathbf{F}$. Hence it can always precode the transmitted data $\mathbf{x}$ to be $\mathbf{Px}$, even when sending the training data. Thus the receiver is "fooled" to believe that the channel is the virtual one $\mathbf{H}_{vt} = \mathbf{HP} = \mathbf{QR}$. Although the linear precoder $\mathbf{P}$ is made transparent to the VBLAST detector, the decoder still enjoys the multiple *identical* subchannels due to the linear precoder $\mathbf{F} = \mathbf{P}$.

3.3.3 Subchannel Selection

The previous discussion is based on the assumption that all the subchannels corresponding to positive singular values are used for signal transmission. However, in practical scenarios, some of the positive singular values of the channel matrix $\mathbf{H}$ can be very small. This situation occurs for spatially correlated flat fading channels, or even i.i.d. Rayleigh flat fading channels with $M_r \approx M_t \gg 1$. From (3.12b), we see that it will influence the overall channel quality and hence subchannel selection is helpful. The other situation where subchannel selection is needed is the case when the input power is low or moderate. In this section, we propose a simple algorithm to select the

subchannels, which is numerically verified to be able to achieve near optimal capacity even at low SNR.

Let us sort the singular values of $\mathbf{H}$ as $\lambda_{H,1} \geq \lambda_{H,2} \ldots \geq \lambda_{H,K} > 0$. If GMD is constrained to the first $n \leq K$ eigen subchannels, we obtain n identical subchannels

$$y_i = \bar{\lambda}_n x_i + z_i, \quad \text{for} \quad i = 1, \ldots, n. \tag{3.29}$$

where

$$\bar{\lambda}_n = \sqrt[n]{\prod_{i=1}^{n} \lambda_{H,i}}. \tag{3.30}$$

To maximize the channel throughput with our GMD scheme, we need to solve the following problem

$$\max_{1 \leq n \leq K} n \log \left(1 + \sqrt[n]{\prod_{i=1}^{n} \lambda_{H,i}^2} \frac{\rho}{n} \right) \tag{3.31}$$

or

$$\max_{1 \leq n \leq K} \left(1 + \sqrt[n]{\prod_{i=1}^{n} \lambda_{H,i}^2} \frac{\rho}{n} \right)^n. \tag{3.32}$$

The solution to this problem is straightforward. We can use either linear search or bisection method to find the optimal n.

Several remarks are in order. i) It is straightforward to incorporate the channel selection into the GMD algorithm. In Section 6.2.2, we show that GMD starts from SVD $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$ and then applies a series of Givens transformation to $\mathbf{\Lambda}$ to make it upper triangular. The Givens transformation can be constrained to the first $n \leq K$ diagonal elements of $\mathbf{\Lambda}$. ii) The blind channel subspace tracking can be combined with the subchannel selection strategy seamlessly. If only the subchannels corresponding to the largest $n < K$ singular values are selected, the blind channel tracking technique will track the n dimensional subspace automatically. iii) The performance loss of the GMD scheme at low SNR region is due to the well-known fact that the zero-forcing equalizer is inherently suboptimal. In the next chapter, we propose the so-called uniform channel decomposition (UCD) scheme, which can decompose a MIMO channel into multiple identical subchannels in a *strictly* capacity lossless manner.

3.3.4 Further Remarks

The author later noticed [30] in which an idea similar to GMD was proposed to approach the performance of the ML detector in the ISI suppression scenario. For a SISO ISI channel, if symbols are precoded and transmitted in a block manner, then the

data model (2.39) can be used to represent the received block data (cf. (2.3) and (2.4)). Note that for this case, $\mathbf{H}$ is a Toeplitz matrix due to the time invariant property of the ISI channel. A linear precoder design $\mathbf{F}$ was proposed in [30] such that the virtual channel $\mathbf{H}_{vt} = \mathbf{HF}$ can be decomposed via QR decomposition to be $\mathbf{H}_{vt} = \mathbf{QR}$ where $\mathbf{R}$ has equal diagonal elements. We see that this equal diagonal idea is equivalent to GMD. However, our GMD scheme, independently motivated by the MIMO transceiver design problem, has several major advantages over the algorithm in [30]:

1. Our GMD scheme represents a paradigm shift from the conventional linear transceiver designs to nonlinear designs and can be proven, both numerically and theoretically, to have superior performance from both BER and information theoretic aspects.
2. Our GMD algorithm is computationally much more efficient than that of [30]. Both algorithms start from the SVD of $\mathbf{H}$ which is followed by $K-1$ iterations. The GMD involves $2K-2$ fast Givens rotations. For a channel $\mathbf{H}$ with $M_t = M_r = K$, the SVD requires $O(K^3)$ flops while the GMD requires additional $O(K^2)$ flops. Thus the computational complexity of the GMD scheme is comparable to the conventional linear transceiver schemes. However, the algorithm in [30] involves multiplications and inversions of matrices in each iteration and the overall computational burden turns out to be additional $O(K^4)$ flops.
3. For the GMD algorithm, only the information of $\mathbf{HH}^*$, and hence $\mathbf{\Lambda}$ and $\mathbf{U}$, are needed to calculate $\mathbf{Q}$. However, for the algorithm in [30], the equalizer needs to know both the precoder $\mathbf{F}$ and $\mathbf{H}$, and hence $\mathbf{H}_{vt} = \mathbf{HF}$, in order to apply the traditional QR to $\mathbf{H}_{vt}$. Hence it cannot be combined with the aforementioned blind two-way channel subspace tracking algorithm introduced in Section 3.3.2.

Like the algorithm in [30], the GMD scheme can also be combined with orthogonal frequency division multiplexing (OFDM) for ISI suppression. For a SISO ISI channel with memory L,

$$y(n) = \sum_{l=0}^{L-1} h_l x(n-l) + z(n), \tag{3.33}$$

after applying OFDM with block length N, we get a MIMO channel

$$\mathbf{y} = \mathbf{Dx} + \mathbf{z} \tag{3.34}$$

where $\mathbf{D}$ is a diagonal matrix with the diagonal elements equal to the N-point FFT of $\mathbf{h} = [h_0, h_1, \ldots, h_{L-1}]^\mathsf{T}$. Hence the GMD scheme can be applied directly. We expect

that GMD-ZFDP may have better BER performance than GMD-VBLAST if $N \gg 1$, in which case the GMD-VBLAST may suffer from considerable performance degradation due to error propagations.

3.4 Performance Examples

We present next several numerical examples to demonstrate the effectiveness of the GMD scheme. In all the examples, we assume Rayleigh independent flat fading channels.

In the first example, we consider a Rayleigh flat fading channel with $M_t = 4$ and $M_r = 4$. We compute the Shannon capacities of the channel with both CSIR and CSIT (C_{IT}, (2.10)), the channel with uninformed transmitter (C_{UT}, (2.11)), the channel using the GMD scheme (C_{GMD}, (3.16)), the channel using the MTM scheme (C_{MTM}, (2.49)), and the channel using the MMD scheme (C_{MMD}, (2.55)). We average the capacities of 1000 Monte-Carlo-generated $\mathbf{H}$ realizations. The result is presented in Figure 3–1. We note that the capacity loss of the MMD scheme is about twice that of the MTM scheme at high SNR as predicted in Section 2.3. The relative capacity loss of the MMD scheme compared with MTM is smaller at low SNR because some subchannels are not used at low SNR. The GMD scheme outperforms the linear transceiver designs when the SNR is moderate or high and is asymptotically capacity lossless at high SNR.

Figure 3–2 shows the complementary cumulative distribution functions (CCDF) of the channel capacities of a 5×5 independent Rayleigh flat fading channel with SNR equal to 23 dB. The five thin dashed curves denote the channel capacities of the five subchannels obtained via SVD plus water filling. Note that the leftmost thin curve crosses the vertical axis at a value less than one, which means that the worst subchannel (corresponding to the smallest singular value of the channel matrix) is sometimes discarded by water filling. The thick line is the CCDF of each subchannel capacity obtained via GMD. Figure 3–2 further illustrates the disadvantages of the conventional "SVD plus bit allocation" scheme (see, e.g., [19] [20] [23]). The channel capacities of the 5 subchannels obtained via SVD plus water filling range from 0 to about 10 bps/Hz, which suggests that the BPSK or QPSK modulation should be used to match the capacity of the worst subchannel and something like 512 or 1024 QAM to the best subchannel. This bit allocation significantly increases the modulation/demodulation complexity. Moreover, using a constellation with size greater than 256 is impractical for the current RF circuit design technology. For the GMD scheme, on the other hand, the same constellation with a moderate size, say 64-QAM, can be applied to reap most of the channel capacity.

To demonstrate the effectiveness of the subchannel selection approach, we consider a 10×10 independent Rayleigh flat fading channel. The channel is usually ill-conditioned since some singular values of $\mathbf{H}$ are very close to zero. Without the subchannel selection strategy, GMD suffers from performance degradation, especially at low SNR, as seen in Figure 3–3. On the other hand, with the subchannel selection scheme, there is only about 0.2 bit/sec/Hz rate loss compared with the C_{IT}, even at very low SNR.

We compare the BER performance of the GMD-VBLAST scheme with the unprecoded MMSE-VBLAST scheme with the optimal detection ordering, the MTM scheme and the MMD scheme. No error correcting code is used in the simulations. In Figure 3–4(a), $\mathbf{H} \in \mathbb{C}^{4\times 2}$ has identically independent Rayleigh fading elements. Hence the channel matrix is usually well-conditioned. Two independent symbol streams modulated as 16-QAM are transmitted. The figure is obtained by averaging 1000 Monte Carlo trials of $\mathbf{H}$. We see that the GMD scheme has more than one dB improvement over the MMD scheme at moderate to high SNR. In Figure 3–4(b), $\mathbf{H} \in \mathbb{C}^{4\times 4}$ usually has a large condition number, in which case the MMD scheme is subject to more capacity loss as analyzed in Section 2.3. Four independent symbol streams are transmitted. The BER performance of the GMD scheme is much better than the others. We did not include MTM because it discards some bad subchannels and hence cannot be used to transmit four independent data streams.

In the final example, we combine the GMD scheme with 64-point FFT based OFDM for ISI suppression in a SISO channel. We assume that the channel response $h_l, l = 0, 1, \ldots, L-1$, are independent zero-mean circularly symmetric Gaussian random variables with unit variance. The channel length is $L = 4$. The GMD-ZFDP is about 2 dB better than GMD-VBLAST. This is because GMD-VBLAST suffers from considerable error propagation effect. This result suggests that GMD-ZFDP may be preferred over GMD-VBLAST if the channel has a large dimensionality.

3.5 Conclusions

In this chapter, we introduce a novel joint transceiver design, which combining the geometric mean decomposition (GMD) with the VBLAST equalizer or dirty paper precoder. The GMD scheme can decompose a MIMO channel into multiple *identical* scalar subchannels. This desirable property can bring about much convenience to the practical system design, particularly the symbol constellation selection. Moreover, we have shown that the GMD scheme is optimal asymptotically for high SNR in terms of

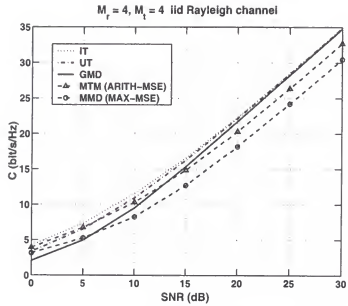

Figure 3–1: Average capacity over 1000 Monte Carlo trials vs. SNR with $M_t = 4$ and $M_r = 4$ for i.i.d. Rayleigh flat fading channels.

both information rate and BER performance while the computational complexity of our scheme is comparable with the conventional linear transceiver scheme. Furthermore, we have shown that the GMD scheme can be applied without the need of using training symbols for channel estimation if combined with subspace tracking techniques. We have also considered the issue of subchannel selection when some of the subchannels are too poor to be useful. The GMD scheme can also be combined with OFDM for ISI suppression. Both the theoretical analyses and empirical simulations have been provided to validate the effectiveness of our approaches.

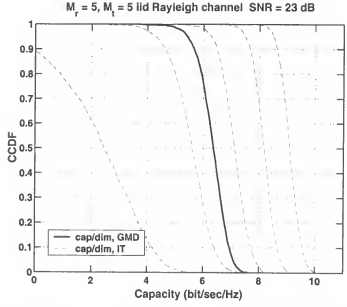

Figure 3–2: Complementary cumulative distribution functions of the capacities of 5 subchannels of the i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 5$. Results based on 2000 Monte Carlo trials.

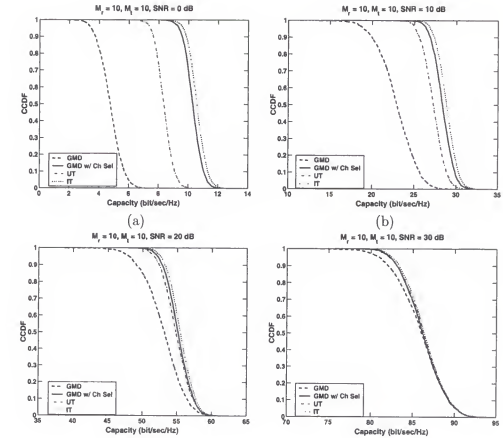

Figure 3–3: Complementary cumulative distribution function of the capacity of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$. Results based on 1000 Monte Carlo trials. SNR = (a) 0 dB, (b) 10 dB, (c) 20 dB, and (d) 30 dB.

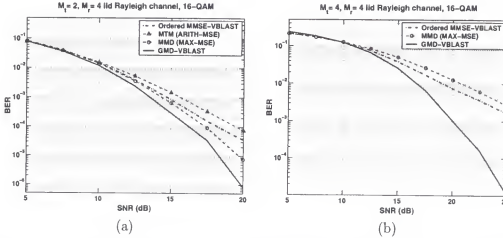

Figure 3–4: BER performance averaged over 1000 Monte Carlo trials of i.i.d. Rayleigh flat fading channel vs. SNR with (a) $M_t = 2$ and $M_r = 4$ and (b) $M_t = 4$ and $M_r = 4$.

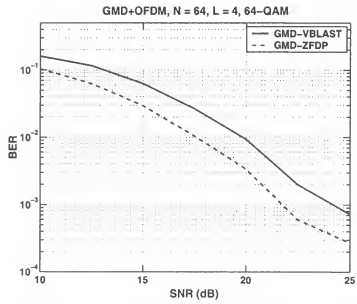

Figure 3–5: BER performances of GMD-VBLAST and GMD-ZFDP. Both are combined with OFDM for ISI suppression.

CHAPTER 4
UNIFORM CHANNEL DECOMPOSITION

We have seen in Chapter 3 that the GMD scheme can have much better performance than the conventional linear transceivers. However, the GMD scheme may suffer from considerable capacity loss at low SNR due to the inherent "zero-forcing" operations which is capacity lossy, especially at low SNR. In this chapter, we propose a uniform channel decomposition (UCD) scheme, which is also based on the GMD matrix decomposition algorithm, to decompose a MIMO channel into multiple *identical* subchannels. The UCD scheme has two implementation forms. One is the combination of a linear precoder and a minimum mean-squared-error VBLAST (MMSE-VBLAST) detector, which is referred to as UCD-VBLAST, and the other includes a dirty paper (DP) precoder and a linear equalizer followed by a DP decoder, which we refer to as UCD-DP. Just like the GMD scheme, UCD can bring much convenience to the subsequent modulation/demodulation and coding/decoding procedures by obviating the need of bit allocation. Two remarkable merits of UCD, which are not shared by the GMD scheme, are that first, UCD is *strictly* capacity lossless at any SNR, and second, UCD has the maximal diversity gain. Moreover, the UCD scheme can decompose a MIMO channel into an arbitrarily large number of independent subchannels, which is an enabling technology to achieve high data rate transmission using small symbol constellations.

To facilitate the discussion, we recall the channel model given in (2.1) as follows.

$$\mathbf{y} = \mathbf{HFx} + \mathbf{z}, \tag{4.1}$$

where $\mathbf{x} \in \mathbb{C}^{L\times 1}$ is the information symbols precoded by the linear precoder $\mathbf{F} \in \mathbb{C}^{M_t\times L}$ and $\mathbf{y} \in \mathbb{C}^{M_r\times 1}$ is the received signal and $\mathbf{H} \in \mathbb{C}^{M_r\times M_t}$ is the channel matrix with rank K. We assume $E[\mathbf{xx}^*] = \sigma_x^2\mathbf{I}_L$ and $\mathbf{z} \sim N(0, \sigma_z^2\mathbf{I}_{M_r})$ is the circularly symmetric complex Gaussian noise. We define the input SNR as

$$\rho = \frac{E[\mathbf{x}^*\mathbf{F}^*\mathbf{Fx}]}{\sigma_z^2} = \frac{\sigma_x^2}{\sigma_z^2}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\} \triangleq \frac{1}{\alpha}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\}. \tag{4.2}$$

4.1 Closed-Form Representation of MMSE-VBLAST

The UCD scheme is based on the closed-form representation of the VBLAST scheme using MMSE nulling vectors. For MMSE-VBLAST, the nulling vector for the ith layer is

$$\mathbf{w}_i = \left(\sum_{j=1}^{i} \mathbf{h}_j \mathbf{h}_j^* + \alpha \mathbf{I} \right)^{-1} \mathbf{h}_i, \quad i = 1, \ldots, M_t. \tag{4.3}$$

The MMSE-VBLAST algorithm can be represented in a concise matrix form which was given in [9] (also see the more detailed version [47]).

Consider the augmented matrix

$$\mathbf{H}_a = \begin{bmatrix} \mathbf{H} \\ \sqrt{\alpha}\mathbf{I}_{M_t} \end{bmatrix}_{(M_r+M_t)\times M_t} \tag{4.4}$$

Applying the QR decomposition to $\mathbf{H}_a$ yields

$$\mathbf{H}_a = \mathbf{Q}_{H_a}\mathbf{R}_{H_a} \triangleq \begin{bmatrix} \mathbf{Q}_{H_a}^u \\ \mathbf{Q}_{H_a}^l \end{bmatrix} \mathbf{R}_{H_a} \tag{4.5}$$

where $\mathbf{R}_{H_a} \in \mathbb{C}^{M_t \times M_t}$ is an upper triangular matrix with positive diagonal elements and $\mathbf{Q}_{H_a}^u \in \mathbb{C}^{M_r \times M_t}$. Note that $\mathbf{H} = \mathbf{Q}_{H_a}^u \mathbf{R}_{H_a}$ is *not* the QR decomposition of $\mathbf{H}$ since $\mathbf{Q}_{H_a}^u$ is not unitary. However, we can readily obtain the nulling vectors using $\mathbf{Q}_{H_a}^u$ and $\mathbf{R}_{H_a}$ as shown in the following lemma [47]:

Lemma 4.1.1 *Let* $\{\mathbf{q}_{H_a,i}\}_{i=1}^{M_t}$ *denote the columns of* $\mathbf{Q}_{H_a}^u$ *and* $\{r_{H_a,ii}\}_{i=1}^{M_t}$ *the diagonal elements of* $\mathbf{R}_{H_a}$*, where* $\mathbf{Q}_{H_a}^u$ *and* $\mathbf{R}_{H_a}$ *are given in (4.5). The nulling vectors of (4.3) satisfy*

$$\mathbf{w}_i = r_{H_a,ii}^{-1}\mathbf{q}_{H_a,i}, \quad i = 1, 2, \ldots, M_t. \tag{4.6}$$

Then the output signal-to-interfere-and-noise ratio (SINR) of the ith layer (i.e., the signal corresponding to $\mathbf{h}_i$) using $\mathbf{w}_i$ is

$$\rho_i = \frac{|\mathbf{h}_i^* \mathbf{w}_i|^2 \sigma_x^2}{\mathbf{w}_i^* \left(\sum_{j=1}^{i-1} \sigma_x^2 \mathbf{h}_j \mathbf{h}_j^* + \sigma_z^2 \mathbf{I} \right) \mathbf{w}_i}. \tag{4.7}$$

Inserting (4.3) into (4.7), we can simplify (4.7) via some straightforward calculations to be (see, e.g., [48])

$$\rho_i = \mathbf{h}_i^* \mathbf{C}_i^{-1} \mathbf{h}_i, \quad i = 1, \ldots, M_t. \tag{4.8}$$

where $\mathbf{C}_i = \sum_{j=1}^{i-1} \mathbf{h}_j \mathbf{h}_j^* + \alpha \mathbf{I}$.

The SINRs given in (4.8) are related to $\mathbf{R}_{H_a}$ as shown in the following lemma:

Lemma 4.1.2 *The diagonal of* $\mathbf{R}_{H_a}$ *given in (4.5) and* $\{\rho_i\}_{i=1}^{M_t}$ *given in (4.8) satisfy*

$$\alpha(1+\rho_i) = r^2_{H_a,ii}, \quad i = 1, 2, \ldots, M_t. \tag{4.9}$$

Proof: See Appendix A. ■

An immediate corollary follows.

Corollary 4.1.3 *The MMSE-VBLAST detector is information lossless. That is,*

$$\sum_{i=1}^{M_t} \log(1+\rho_i) = \log|\mathbf{H}^*\mathbf{H}\alpha^{-1} + \mathbf{I}|, \tag{4.10}$$

where the right hand side of (4.10) is equal to (2.7) with $\mathbf{F} = \mathbf{I}_{M_t}$.

Proof: From (4.4) and (4.5), we have

$$\mathbf{H}^*\mathbf{H}\alpha^{-1} + \mathbf{I} = \alpha^{-1}\mathbf{H}_a{}^*\mathbf{H}_a = \alpha^{-1}\mathbf{R}_{H_a}{}^*\mathbf{R}_{H_a}. \tag{4.11}$$

Hence

$$\log|\mathbf{H}^*\mathbf{H}\alpha^{-1} + \mathbf{I}| = \sum_{i=1}^{M_t} \log\left(\alpha^{-1} r^2_{H_a,ii}\right). \tag{4.12}$$

According to Lemma 4.1.2,

$$\log|\mathbf{H}^*\mathbf{H}\alpha^{-1} + \mathbf{I}| = \sum_{i=1}^{M_t} \log(1+\rho_i).$$

■

We note that Corollary 4.1.3 coincides with the findings in [48].

4.2 UCD-VBLAST

If we modify the precoder $\mathbf{F}$ given in (2.8) to be

$$\mathbf{F} = \mathbf{V}\Phi^{1/2}\Omega^* \tag{4.13}$$

where $\Omega \in \mathbb{C}^{L\times K}$ with $L \geq K$ (to avoid capacity loss, we should not choose $L < K$ in general) and $\Omega^*\Omega = \mathbf{I}$, then we see through inserting (4.13) into (2.7) that the $\mathbf{F}$ given in (4.13) is also a precoder maximizing the channel throughput. However, introducing Ω brings much greater flexibility than the precoder of (2.8). In the following, we concentrate on how to design Ω.

Given the precoder of (4.13), the virtual channel is

$$\mathbf{G} \triangleq \mathbf{HF} = \mathbf{U}\Lambda\Phi^{1/2}\Omega^* \triangleq \mathbf{U}\Sigma\Omega^* \tag{4.14}$$

where $\mathbf{\Sigma} = \Lambda\Phi^{1/2}$ is a diagonal matrix with diagonal elements $\{\sigma_i\}_{i=1}^{K}$. Let $\mathbf{G}_a$ denote the augmented matrix

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{U}\Sigma\Omega^* \\ \sqrt{\alpha}\mathbf{I}_L \end{bmatrix} \tag{4.15}$$

The UCD scheme is based on the following lemma.

Lemma 4.2.1 *For any matrix of the form given in (4.15), we can find a semi-unitary matrix $\Omega \in \mathbb{C}^{L\times K}$ such that the QR decomposition of $\mathbf{G}_a$ yields an upper triangular matrix with equal diagonal elements.*

Proof: Rewrite (4.15) as

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{U}[\mathbf{\Sigma} \vdots 0_{K\times(L-K)}]\mathbf{\Omega}_0^* \\ \sqrt{\alpha}\mathbf{I}_L \end{bmatrix} \tag{4.16}$$

where $\mathbf{\Omega}_0 \in \mathbb{C}^{L\times L}$ is a unitary matrix whose first K columns form $\mathbf{\Omega}$. We further rewrite (4.16) as

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{I}_{M_r} & 0 \\ 0 & \mathbf{\Omega}_0 \end{bmatrix} \begin{bmatrix} \mathbf{U}[\mathbf{\Sigma} \vdots 0_{K\times(L-K)}] \\ \sqrt{\alpha}\mathbf{I}_L \end{bmatrix} \mathbf{\Omega}_0^*. \tag{4.17}$$

We can have the following GMD:

$$\mathbf{J} \triangleq \begin{bmatrix} \mathbf{U}[\mathbf{\Sigma} \vdots 0_{K\times(L-K)}] \\ \sqrt{\alpha}\mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J\mathbf{R}_J\mathbf{P}_J^* \tag{4.18}$$

where $\mathbf{R}_J \in \mathbb{R}^{L\times L}$ is an upper triangular matrix with equal diagonal elements and $\mathbf{Q}_J \in \mathbb{C}^{(M_r+L)\times L}$ is semi-unitary and $\mathbf{P}_J \in \mathbb{C}^{L\times L}$ is unitary. Inserting (4.18) into (4.17) yields

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{I}_{M_r} & 0 \\ 0 & \mathbf{\Omega}_0 \end{bmatrix} \mathbf{Q}_J\mathbf{R}_J\mathbf{P}_J^*\mathbf{\Omega}_0^*. \tag{4.19}$$

Let $\mathbf{\Omega}_0 = \mathbf{P}_J^*$ and

$$\mathbf{Q}_{G_a} = \begin{bmatrix} \mathbf{I}_{M_r} & 0 \\ 0 & \mathbf{\Omega}_0^* \end{bmatrix} \mathbf{Q}_J. \tag{4.20}$$

Then (4.19) can be rewritten to be $\mathbf{G}_a = \mathbf{Q}_{G_a}\mathbf{R}_J$ which is the QR decomposition of $\mathbf{G}_a$. The semi-unitary matrix $\mathbf{\Omega}$ associated with $\mathbf{G}_a$ consists of the first K columns of $\mathbf{\Omega}_0$ (or $\mathbf{P}_J^*$). ■

From Lemma 4.2.1 and Lemma 4.1.2, we conclude that we can always combine a linear precoder and the MMSE-VBLAST detector to uniformly decompose a MIMO channel into $L \geq K$ subchannels with the same output SINRs. According to Corollary 4.1.3, we can further conclude that the channel decomposition is strictly capacity lossless. We refer to the scheme demonstrated in Lemma 4.2.1 as UCD-VBLAST.

The proof of Lemma 4.2.1 is insightful. Indeed, given the SVD of $\mathbf{H}$ and the "water filling" level $\mathbf{\Phi}^{1/2}$, we only need to calculate the GMD given in (4.18). Then we immediately obtain the linear precoder $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\star}$, where $\mathbf{\Omega}$ consists of the first K columns of $\mathbf{P}_J^{\star}$. Let $\mathbf{Q}_{G_a}^u$ denote the first M_r rows of $\mathbf{Q}_{G_a}$, or equivalently the first M_r rows of $\mathbf{Q}_J$ (cf. (4.20)). According to Lemma 4.1.1, the nulling vectors are calculated as

$$\mathbf{w}_i = r_{J,ii}^{-1}\mathbf{q}_{G_a,i}, \quad i = 1, 2, \ldots, L \tag{4.21}$$

where $r_{J,ii}$ is the ith diagonal element of $\mathbf{R}_J$ and $\mathbf{q}_{G_a,i}$ is the ith column of $\mathbf{Q}_{G_a}^u$.

Some observations can help reduce the computational complexity. For any matrix $\mathbf{B} \in \mathbb{C}^{M\times N}$ with SVD $\mathbf{B} = \mathbf{U}_B\mathbf{\Lambda}_B\mathbf{V}_B^{\star}$ and the augmented matrix with SVD

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} \\ \sqrt{\alpha}\mathbf{I} \end{bmatrix} = \mathbf{U}_A\mathbf{\Lambda}_A\mathbf{V}_A^{\star}, \tag{4.22}$$

the diagonal elements of $\mathbf{\Lambda}_A$ and $\mathbf{\Lambda}_B$, i.e., $\lambda_{A,i}$ and $\lambda_{B,i}$, satisfy

$$\lambda_{A,i} = \sqrt{\lambda_{B,i}^2 + \alpha}, \quad i = 1, \ldots, N. \tag{4.23}$$

Moreover

$$\mathbf{U}_A = \begin{pmatrix} \mathbf{U}_B\mathbf{\Lambda}_B\mathbf{\Lambda}_A^{-1} \\ \sqrt{\alpha}\mathbf{V}_B^{\star}\mathbf{\Lambda}_A^{-1} \end{pmatrix} \quad \text{and} \quad \mathbf{V}_A = \mathbf{V}_B. \tag{4.24}$$

Hence the SVD of $\mathbf{J}$ defined in (4.18) is

$$\mathbf{J} = \begin{bmatrix} \mathbf{U}[\mathbf{\Sigma} \vdots 0_{K\times(L-K)}]\tilde{\mathbf{\Sigma}}^{-1} \\ \sqrt{\alpha}\tilde{\mathbf{\Sigma}}^{-1} \end{bmatrix} \tilde{\mathbf{\Sigma}}\mathbf{I}_L \tag{4.25}$$

where $\tilde{\mathbf{\Sigma}}$ is an $L \times L$ diagonal matrix with the diagonal elements

$$\tilde{\sigma}_i = \sqrt{\sigma_i^2 + \alpha}, \quad 1 \leq i \leq K, \tag{4.26a}$$

and

$$\tilde{\sigma}_i = \sqrt{\alpha}, \quad K+1 \leq i \leq L. \tag{4.26b}$$

Applying the GMD matrix decomposition algorithm given in Section 6.2 to $\tilde{\mathbf{\Sigma}}$ yields

$$\tilde{\mathbf{\Sigma}} = (\mathbf{Q}_1\mathbf{Q}_2\dots\mathbf{Q}_{L-1})\mathbf{R}_J(\mathbf{P}_{L-1}^{\mathsf{T}}\mathbf{P}_{L-2}^{\mathsf{T}}\dots\mathbf{P}_1^{\mathsf{T}}). \tag{4.27}$$

Hence

$$\begin{bmatrix} \mathbf{U}[\mathbf{\Sigma}\,\vdots\,\mathbf{0}_{K\times(L-K)}] \\ \sqrt{\alpha}\mathbf{I}_L \end{bmatrix} = \begin{bmatrix} \mathbf{U}[\mathbf{\Sigma}\,\vdots\,\mathbf{0}_{K\times(L-K)}]\tilde{\mathbf{\Sigma}}^{-1} \\ \sqrt{\alpha}\tilde{\mathbf{\Sigma}}^{-1} \end{bmatrix} (\mathbf{Q}_1\mathbf{Q}_2\dots\mathbf{Q}_{L-1})\mathbf{R}_J(\mathbf{P}_{L-1}^{\mathsf{T}}\mathbf{P}_{L-2}^{\mathsf{T}}\dots\mathbf{P}_1^{\mathsf{T}}). \tag{4.28}$$

Then the linear precoder has the form:

$$\mathbf{F} = \mathbf{V}\left[\mathbf{\Phi}^{1/2}\,\vdots\,\mathbf{0}_{K\times(L-K)}\right]\mathbf{P}_1\mathbf{P}_2\dots\mathbf{P}_{L-1}. \tag{4.29}$$

The nulling vectors are calculated according to (4.21) with $r_{J,ii} = \left(\prod_{i=1}^{L}\tilde{\sigma}_i\right)^{1/L}$, and

$$\mathbf{Q}_{G_a}^{u} = \mathbf{U}[\mathbf{\Sigma}\,\vdots\,\mathbf{0}_{K\times(L-K)}]\tilde{\mathbf{\Sigma}}\mathbf{Q}_1\mathbf{Q}_2\dots\mathbf{Q}_{L-1}. \tag{4.30}$$

Note that $\mathbf{Q}_l$ and $\mathbf{P}_l, l = 1, 2, \dots, L$, are Givens rotation matrices and hence calculating (4.29) and (4.30) needs $O(M_t(K+L))$ and $O(M_r(K+L))$ flops, respectively.

We summarize the UCD-VBLAST scheme as follows [1]

Table 4-1: The UCD-VBLAST scheme

step	operation	flops
1	Compute SVD $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$	$O(M_tM_rK)$
2	Calculate $\mathbf{\Phi}^{1/2}$ using (2.9)	$O(K^2)$
3	$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}^{1/2}$	$O(K)$
4	Obtain $\tilde{\mathbf{\Sigma}}$ using (4.26)	$O(K)$
5	Apply GMD to $\tilde{\mathbf{\Sigma}}$ to obtain (4.27)	$O(L^2)$
6	Generate $\mathbf{F}$ using (4.29)	$O(M_t(K+L))$
7	Compute $\mathbf{Q}_{G_a}^{u}$ using (4.30)	$O(M_r(K+L))$
8	Calculate $\{\mathbf{w}_i\}_{i=1}^{L}$ using (4.21)	$O(M_rL)$

Obviously, our UCD-VBLAST scheme has comparable computational complexity to the SVD based linear transceiver designs. An observation relevant to practical implementations is as follows. Note that the receiver does not have to calculate Step 6 since CSIT is available and the transmitter can run Steps 1 to 6. However, if the receiver calculates $\mathbf{F}$, which only takes a small number of flops, and feeds it back to the transmitter,

[1] Steps 5-7 can be processed simultaneously as in the GMD algorithm.

then the transmitter is relieved from calculating the SVDs. Hence in FDD systems, it is preferable to feed back $\mathbf{F}$, rather than $\mathbf{H}$, to the transmitter. In TDD systems, there are still advantages for feeding back $\mathbf{F}$ since this reduces by approximately half the overall computational complexity.

We conclude the discussions of the UCD-VBLAST scheme by deriving the SINR of each subchannel. Note that the diagonal elements of $\mathbf{R}_J$ is

$$r_{J,ii} = \left(\prod_{l=1}^{L} \tilde{\sigma}_l\right)^{1/L}, \quad i = 1, 2, \ldots, L, \tag{4.31}$$

which is the geometric mean of the diagonal elements of $\tilde{\mathbf{\Sigma}}$. It follows from (4.26) that

$$r_{J,ii}^2 = \left(\alpha^{L-K} \prod_{l=1}^{K} (\sigma_l^2 + \alpha)\right)^{1/L} = \alpha \left(\prod_{l=1}^{K} (\alpha^{-1}\sigma_l^2 + 1)\right)^{1/L}. \tag{4.32}$$

According to Lemma 4.1.2,

$$\rho_i = \bar{\rho} \triangleq \left(\prod_{l=1}^{K} (\alpha^{-1}\sigma_l^2 + 1)\right)^{1/L} - 1, \quad i = 1, 2, \ldots, L. \tag{4.33}$$

Hence

$$\sum_{i=1}^{L} \log_2(1 + \rho_i) = \sum_{i=1}^{K} \log_2(1 + \alpha^{-1}\sigma_i^2) = \sum_{i=1}^{K} \log_2(1 + \alpha^{-1}\lambda_{H,i}^2 \phi_i) \tag{4.34}$$

which is exactly the C_{IT} in (2.10). Hence UCD-VBLAST is strictly capacity lossless.

4.3 UCD-DP

As a dual form of UCD-VBLAST, the UCD scheme can be implemented by using DP precoding, which we refer to as UCD-DP. For UCD-DP, a direct construction of the linear precoder $\mathbf{F}$ as done in Section 4.2 is not obvious. Instead, we exploit the uplink-downlink duality revealed in [49] to obtain UCD-DP.

We convert the UCD-DP problem into the UCD-VBLAST problem in the reverse channel where the roles of the transmitter and receiver are exchanged

$$\mathbf{y} = \mathbf{H}^* \mathbf{x} + \mathbf{z}. \tag{4.35}$$

The UCD-VBLAST scheme can be applied to the channel of (4.35), which yields the precoder $\mathbf{F}_{rev}$ and the equalizer $\{\mathbf{w}_i\}_{i=1}^L$ as in (4.29) and (4.21), respectively. Normalize $\{\mathbf{w}_i\}_{i=1}^L$ to be of unit Euclidean norm, which we denote as $\{\bar{\mathbf{w}}_i\}_{i=1}^L$. Let $\mathbf{W} = [\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_L]$. According to the uplink-downlink duality, the precoder of UCD-DP should be $\mathbf{F} = \mathbf{W}\mathbf{D}_q$ where $\mathbf{D}_q$ is diagonal with the diagonal elements $\{\sqrt{q_i}\}_{i=1}^L$, which will be

determined based on (4.40) below. We use $\mathbf{F}_{rev}$, the linear precoder in the reverse channel, as the linear equalizer. Then the equivalent MIMO channel is

$$\mathbf{y} = \mathbf{F}_{rev}^{*}\mathbf{H}\mathbf{W}\mathbf{D}_{q}\mathbf{x} + \mathbf{F}_{rev}^{*}\mathbf{z}, \tag{4.36}$$

where the ith scalar subchannel of the MIMO channel is

$$y_i = \mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_i\sqrt{q_i}x_i + \sum_{j=i+1}^{L}\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j\sqrt{q_j}x_j + \sum_{j=1}^{i-1}\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j\sqrt{q_j}x_j + \mathbf{f}_i^{*}\mathbf{z}. \tag{4.37}$$

Applying the dirty paper precoder to x_i and treating $\sum_{j=1}^{i-1}\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j\sqrt{q_j}x_j$ as the interference known at the transmitter (note that here we precode the first layer first while for UCD-VBLAST, we detect the Lth layer first), we obtain an equivalent subchannel

$$y_i = \mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_i\sqrt{q_i}x_i + \sum_{j=i+1}^{L}\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j\sqrt{q_j}x_j + \mathbf{f}_i^{*}\mathbf{z} \tag{4.38}$$

with SINR

$$\rho_i = \frac{q_i|\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_i|^2}{\alpha\|\mathbf{f}_i\|^2 + \sum_{j=i+1}^{L}q_j|\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j|^2} \quad \text{for } i = 1, 2, \ldots, L. \tag{4.39}$$

The next step is to calculate $\{q_i\}_{i=1}^{L}$ such that $\rho_i = \bar{\rho},\ 1 \le i \le L$, where $\bar{\rho}$ is as defined in (4.33). Let $a_{ij} = |\mathbf{f}_i^{*}\mathbf{H}\bar{\mathbf{w}}_j|^2$. Then (4.39) can be represented in the matrix form

$$\begin{bmatrix} a_{11} & -\bar{\rho}a_{12} & \cdots & -\bar{\rho}a_{1L} \\ 0 & a_{22} & \cdots & -\bar{\rho}a_{2L} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{LL} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_L \end{bmatrix} = \bar{\rho}\alpha \begin{bmatrix} \|\mathbf{f}_1\|^2 \\ \|\mathbf{f}_2\|^2 \\ \vdots \\ \|\mathbf{f}_L\|^2 \end{bmatrix}. \tag{4.40}$$

It is easy to see that $q_i > 0,\ 0 \le i \le L$. It is proven in [49] that $\sum_{i=1}^{L} q_i = \mathrm{tr}(\mathbf{F}\mathbf{F}^{*}) = \mathrm{tr}(\mathbf{F}_{rev}\mathbf{F}_{rev}^{*})$. That is, the UCD-DP needs exactly the same power as the UCD-VBLAST to obtain L identical subchannels with SINR $\bar{\rho}$.

The UCD-DP using the Tomlinson-Harashima precoder leads to an input power increase of $\frac{M}{M-1}$ for M-QAM symbols. Nevertheless, for a system with high dimensionality and/or using large constellation, UCD-DP is a better choice than UCD-VBLAST since it is free of propagation errors.

4.4 Performance Analysis

4.4.1 Diversity Gain Analysis

An important performance metric is diversity gain, which is defined as follows [16].

Definition 4.4.1 *Let $P_e(\rho)$ denote the average error probability of a scheme at SNR ρ. The diversity gain of the scheme is*

$$d = -\lim_{\rho \to \infty} \frac{\log P_e(\rho)}{\log \rho}. \tag{4.41}$$

The diversity gain measures how fast the error probability decays with SNR. We note that diversity gain is usually discussed without assuming the availability of CSIT. The reason is that diversity gain is a concept associated with channel outage, i.e., the case where the channel is too poor to support a target data rate. Using CSIT, one can adjust the transmission rate to avoid channel outage. However, if the rate is fixed, which is desirable in practice, we can also use diversity gain as a performance measure of the transceiver designs. Based on this observation, we analyze the diversity gains of the UCD and GMD schemes. The result is summarized in the following proposition.

Proposition 4.4.2 *Consider the i.i.d. Rayleigh flat fading MIMO channel defined in (4.1). Let $M = \max(M_t, M_r)$ and $m = \min(M_t, M_r)$. The diversity gains of the GMD and the UCD schemes are*

$$d_{GMD}(M,m) = (M - m + 1)m, \quad \textit{and} \quad d_{UCD}(M,m) = Mm, \tag{4.42}$$

respectively.

We have applied the *typical error event* analysis (see [16][50]) to obtain (4.42). The details are relegated to Appendix B. We see that although UCD has a negligible coding gain compared with the GMD scheme at high SNR, it has an additional $m^2 - m$ diversity gains over GMD. An interesting point to make is that water filling does *not* help improve diversity gains. Hence at high SNR, water filling is useless in both capacity and diversity aspects.

Given the fact that the GMD scheme is asymptotically capacity lossless for high SNR, it is rather surprising to see the large diversity loss of GMD compared with UCD. We give an intuitive explanation as follows. Note that diversity gain is determined by the typical error events that the MIMO channel is in deep fade. Namely, the diversity gain of a scheme depends on its ability of dealing with bad channels. A deeply faded channel with high input SNR is equivalent to a "normal" channel with low SNR, in which scenario the GMD scheme is far less efficient than UCD as shown in the numerical examples. Consequently, the GMD has less diversity gain than UCD.

4.4.2 Further Remarks

Besides the larger coding gain at low SNR and an improved diversity gain at high SNR, the UCD scheme enjoys more flexibility than the GMD scheme. For a rank K MIMO channel, the GMD scheme can support no more than K independent data streams. However, the UCD scheme can decompose a rank K MIMO channel into $L \geq K$ *identical* subchannels, and L is *not* even limited by the dimensionality of the channel matrix. This property of the UCD scheme enables one to achieve high data rate transmission using small constellations as demonstrated in the numerical examples.

The UCD scheme also suggests new ways of channel decomposition which are much more flexible than the conventional SVD based ones. Indeed, one may chose the permutation matrices and Givens rotations to achieve a wide variety of channel decompositions with some prescribed SINRs as suggested by the generalized triangular decomposition (GTD) (See Chapter 6, [51] [52]). This idea is developed in Chapter 5.

Finally, we link UCD with DBLAST [2], which has been shown to be able to achieve the optimal tradeoff between the channel diversity and multiplexing [16]. We observe that each diagonal layer of DBLAST can be viewed as the interleaving of the vertical layers of VBLAST in the space-time domain and each diagonal layer can be regarded as a virtual subchannel with the same capacity. However, DBLAST requires short and powerful error correcting coding to make the *virtual* subchannel work as a "real" one. This is a major difficulty for the implementation of DBLAST. In addition, DBLAST suffers from boundary wastage. In contrast, our UCD scheme, by exploiting CSIT, applies interleaving (via the Givens rotations and permutations) in the space domain only. This makes the UCD scheme free from the boundary wastage. Moreover, the UCD scheme is decoupled from coding procedures. Indeed, UCD can be concatenated with any error correcting code. Furthermore, UCD makes it easier to design the coding scheme since UCD decomposes a MIMO channel into multiple subchannels with *identical* capacities. Thus in a slowly time varying channel, UCD is much easier to implement than DBLAST despite their duality. This manifests clearly the values of CSIT.

4.5 Numerical Examples

We present next several numerical examples to demonstrate the effectiveness of the UCD scheme.

In the first example, we assume Rayleigh independent flat fading channels with $M_t = 10$ and $M_r = 10$. We compare the channel capacity using the UCD and GMD

schemes. The complementary cumulative distribution functions (CCDF) of the capacity drawn out of 2000 Monte-Carlo realizations of $\mathbf{H}$ are shown in Figure 4–1. We see that the UCD scheme outperforms the GMD scheme significantly at low SNR although the difference becomes smaller at higher SNR.

Figure 4–2 shows the CCDFs of the channel capacities of a 5×5 independent Rayleigh flat fading channel with SNR equal to 25 dB. The five thin dashed curves denote the channel capacities of the five subchannels obtained via SVD plus water filling. Note that the leftmost thin dashed curve crosses the vertical axis at a value less than one, which means that the worst subchannel (corresponding to the smallest singular value of the channel matrix) is sometimes discarded by water filling. The thick solid line is the CCDF of the capacity of the $L = 5$ subchannels obtained via UCD. All these subchannels have the same capacity. As discussed in Section 4.2, a rank K MIMO channel can be decomposed into $L \geq K$ subchannels. The thin solid line represents the case where a MIMO channel is decomposed into 7 identical subchannels using the UCD scheme. Figure 4–2 demonstrates the advantages of our UCD scheme over the conventional "SVD plus bit allocation" scheme (see, e.g., [19]). The channel capacities of the 5 subchannels obtained via SVD plus water filling range from 0 to about 11 bps/Hz, which suggests that the BPSK or QPSK modulation should be used to match the capacity of the worst subchannel and something like 1024 or 2048 QAM to the best subchannel. This bit allocation significantly increases the modulation/demodulation complexity. Using GMD or UCD, we can decompose a rank 5 MIMO channel into 5 subchannels and hence the same constellation with a reasonable size, say 128-QAM, can be used to reap most of the channel capacity. The UCD scheme can do even better. In this example, after decomposing a MIMO channel into 7 subchannels via UCD, we can apply a small to moderate constellation, say 16-QAM or 64-QAM, to achieve the channel capacity.

In the third example, we assume Rayleigh independent flat fading channels with $M_t = 4$ and $M_r = 4$. We compare the BER performance of the GMD and UCD schemes along with the conventional MMSE-VBLAST with optimal detection ordering in Figure 4–3. We see that both GMD and UCD outperform the conventional VBLAST detector significantly. Moreover, the BER vs. SNR lines of the GMD and UCD schemes have much steeper decreasing slopes, which means much better diversity gains, than the conventional VBLAST. The diversity gains of the GMD and UCD schemes are 4 and 16,

respectively. While there is a noticeably larger diversity gain for UCD compared with GMD as shown in Figure 4–3, the difference is not as drastic as the theoretical prediction. It is because the input SNR is not high enough to validate the approximations made in the typical error event analyses (see Appendix B).

In the final example, we compare the BER performance of UCD-VBLAST and UCD-DP in the scenario of a 10×10 Rayleigh flat fading channel. To present a benchmark, we also include UCD-genie as the imaginary scenario where at each layer, a genie would eliminate the influence of erroneous detections from the previous layers when using UCD-VBLAST. Figure 4–4 shows that UCD-VBLAST may suffer from some small BER degradations caused by error propagation (about 0.5 dB for BER $= 10^{-4}$) compared with UCD-genie. The UCD-DP, on the contrary, is free of error propagation and hence has BER performance very close to that of UCD-genie. The slight SNR loss of UCD-DP is mainly due to the inherent power-amplification effect of the Tomlinson-Harashima precoder.

4.6 Conclusions

Based on the GMD matrix decomposition algorithm and the closed-form representation of the MMSE-VBLAST detector, we have introduced the UCD scheme for MIMO communications that can decompose a MIMO channel into multiple subchannels with *identical* capacities in a capacity lossless manner. We have proposed two versions of the UCD scheme, i.e., UCD-VBLAST and UCD-DP. The UCD scheme can provide much convenience for the subsequent modulation/demodulation and coding/decoding procedures due to obviating the need of bit allocation. We have also shown that UCD can achieve the maximal diversity gain. The simulations show that the UCD scheme has excellent performance even without the use of error correcting codes. The UCD scheme suggests a new way of channel decomposition which enjoys much more flexibility than the conventional SVD based ones.

Appendix A
Proof of Lemma 4.1.2

Rewrite (4.5)

$$\mathbf{H}_a = \mathbf{Q}_{H_a}\mathbf{R}_{H_a} \triangleq \begin{bmatrix} \mathbf{Q}_{H_a}^u \\ \mathbf{Q}_{H_a}^l \end{bmatrix} \mathbf{R}_{H_a}. \tag{4.43}$$

Let $\mathbf{H}_{a,i}$ ($\mathbf{H}_i$) denote the submatrix containing the first i columns of $\mathbf{H}_a$ ($\mathbf{H}$) and $\mathbf{h}_{a,i}$ ($\mathbf{h}_i$) the ith column. Then

$$\mathbf{H}_{a,i} = \begin{bmatrix} \mathbf{H}_i \\ \sqrt{\alpha}\mathbf{I}_i \\ \mathbf{0}_{(M_t-i)\times M_t} \end{bmatrix}, \qquad \mathbf{h}_{a,i} = \begin{bmatrix} \mathbf{h}_i \\ \mathbf{0}_{(i-1)\times 1} \\ \sqrt{\alpha} \\ \mathbf{0}_{(M_t-i)\times 1} \end{bmatrix}. \tag{4.44}$$

For the QR decomposition $\mathbf{H}_a = \mathbf{Q}_{H_a}\mathbf{R}_{H_a}$, the geometric implication of $r_{H_a,ii}$ is the component of $\mathbf{h}_{a,i}$ projected onto the subspace spanned by the ith column of $\mathbf{Q}_{H_a}$, i.e., $\mathbf{q}_{H_a,i}$. Note that $\mathbf{q}_{H_a,i}$ is orthogonal to the subspace spanned by $\{\mathbf{q}_{H_a,j}\}_{j=1}^{i-1}$ or, equivalently, the column space of $\mathbf{H}_{a,i-1}$. Hence

$$r_{H_a,ii}^2 = \mathbf{h}_{a,i}^* \mathcal{P}_{\mathbf{H}_{a,i-1}}^{\perp} \mathbf{h}_{a,i}, \tag{4.45}$$

where $\mathcal{P}_{\mathbf{A}}^{\perp}$ stands for the orthogonal projection onto th null space of $\mathbf{A}^{\mathrm{T}}$. Therefore

$$r_{H_a,ii}^2 = \mathbf{h}_{a,i}^* \left[\mathbf{I} - \mathbf{H}_{a,i-1}\left(\mathbf{H}_{a,i-1}^*\mathbf{H}_{a,i-1}\right)^{-1}\mathbf{H}_{a,i-1}^*\right]\mathbf{h}_{a,i}. \tag{4.46}$$

Inserting (4.44) into (4.46) yields

$$\begin{aligned} r_{H_a,ii}^2 &= \alpha + \mathbf{h}_i^*\left[\mathbf{I} - \mathbf{H}_{i-1}\left(\mathbf{H}_{i-1}^*\mathbf{H}_{i-1} + \alpha\mathbf{I}\right)^{-1}\mathbf{H}_{i-1}^*\right]\mathbf{h}_i \\ &= \alpha + \alpha\mathbf{h}_i^*\left(\mathbf{H}_{i-1}\mathbf{H}_{i-1}^* + \alpha\mathbf{I}\right)^{-1}\mathbf{h}_i. \end{aligned} \tag{4.47}$$

From (4.8), we see that

$$\rho_i = \mathbf{h}_i^*\left(\mathbf{H}_{i-1}\mathbf{H}_{i-1}^* + \alpha\mathbf{I}\right)^{-1}\mathbf{h}_i. \tag{4.48}$$

Hence $r_{H_a,ii}^2 = \alpha(1+\rho_i)$. The lemma is proven.

Appendix B
Proof of Proposition 4.4.2

Without loss of generality, we assume $\mathbf{H} \in \mathbb{C}^{M\times m}$, each of whose entry is of circularly symmetric Gaussian distribution with zero-mean and unit variance. Consider BPSK modulation. The average error probability of the GMD scheme is

$$P_e^{\mathrm{GMD}} = E\left[Q\left(\sqrt{2\rho_{\mathrm{GMD}}}\right)\right] = E\left[Q\left(\sqrt{2\rho\bar{\lambda}_H^2}\right)\right] = E\left[Q\left(\sqrt{2\rho\left(\prod_{i=1}^{m}\lambda_{H,i}^2\right)^{1/m}}\right)\right], \tag{4.49}$$

where the Q-function is defined as

$$Q(x) = \int_x^{+\infty} \frac{1}{2\pi} e^{-\frac{z^2}{2}} dx.$$

The diversity gain of the GMD scheme is

$$d_{\text{GMD}}(M,m) = -\lim_{\rho\to\infty} \frac{\log P_e^{\text{GMD}}}{\log \rho}. \tag{4.50}$$

For any QAM constellation, the average error probability is similar to (4.49) except for some constants before or inside the Q-function. Since we focus on the high SNR region, all these constants will not affect the diversity gain defined in (4.50).

At high SNR, the *typical error event* is

$$\mathcal{E} = \left\{ \bar{\lambda}_H : \bar{\lambda}_H^2 < \rho^{-1} \right\}. \tag{4.51}$$

It can be shown that instead of calculating (4.50), which involves complicated integrations, we can compute the following [50, Ch. 3]:

$$d_{\text{GMD}}(M,m) = -\lim_{\rho\to\infty} \frac{\log P(\mathcal{E})}{\log \rho}. \tag{4.52}$$

Note that

$$\bar{\lambda}_H^{2m} = \prod_{i=1}^{m} \lambda_{H,i}^2 = |\mathbf{H}^*\mathbf{H}|. \tag{4.53}$$

According to [53, Theorem 7.5.3] (with straightforward extensions from real-valued domain to the complex-valued domain),

$$\bar{\lambda}_H^{2m} = |\mathbf{H}^*\mathbf{H}| = \prod_{i=1}^{m} g_{M-m+i}^2 \tag{4.54}$$

where g_i^2's are independent Chi-squared random variables with probability density

$$f_{g_i^2}(x) = \frac{1}{(i-1)!} x^{i-1} e^{-x}, \quad x \geq 0. \tag{4.55}$$

Now the typical error event can be written as

$$\begin{aligned} \mathcal{E} &= \left\{ \{g_{M-m+i}^2\}_{i=1}^m : \prod_{i=1}^{m} g_{M-m+i}^2 < \rho^{-m} \right\} \\ &= \bigcup_{\{\alpha_i\}_{i=1}^m \in \mathcal{E}_\alpha} \left\{ \{g_{M-m+i}^2\}_{i=1}^m : g_{M-m+i}^2 < \rho^{-\alpha_i}, i = 1, \ldots, m \right\}, \end{aligned} \tag{4.56}$$

where $\mathcal{E}_\alpha = \{\{\alpha_i\}_{i=1}^m : \sum_{i=1}^m \alpha_i > m\}$. Hence

$$P(\mathcal{E}) = \int_{\mathcal{E}_\alpha} \prod_{i=1}^{m} P(g_{M-m+i}^2 < \rho^{-\alpha_i}) d\alpha_1 \dots d\alpha_m \tag{4.57}$$

From (4.55), we know that as $\epsilon \to 0$,

$$P(g_i^2 < \epsilon) = \int_0^\epsilon \frac{1}{(i-1)!} x^{i-1} e^{-x} dx \approx \int_0^\epsilon \frac{1}{(i-1)!} x^{i-1} dx = \frac{1}{i!} \epsilon^i. \tag{4.58}$$

Using (4.52) (4.58) and (4.57), we calculate the diversity gain as

$$\begin{aligned} d_{\text{GMD}}(M,m) &= -\lim_{\rho\to\infty} \frac{\log \int_{\mathcal{E}_\alpha^+} \prod_{i=1}^m \frac{\rho^{-(M-m+i)\alpha_i}}{(M-m+i)!} d\alpha_1 \dots d\alpha_m.}{\log \rho} \\ &= -\lim_{\rho\to\infty} \frac{\log \int_{\mathcal{E}_\alpha^+} \rho^{-\sum_{i=1}^m (M-m+i)\alpha_i} d\alpha_1 \dots d\alpha_m.}{\log \rho} \qquad (4.59) \\ &= \inf_{\mathcal{E}_\alpha^+} \sum_{i=1}^{m} (M-m+i)\alpha_i, \qquad (4.60) \end{aligned}$$

where $\mathcal{E}_\alpha^+ = \mathcal{E}_\alpha \bigcap \{\alpha_i > 0, i = 1, \dots, m\}$. To obtain (4.60) from (4.59), we have used the property that the integral in the numerator of (4.59) is dominated by the term with the SNR exponent closest to zero, as $\rho \to \infty$ (see [16] for details). Here the integration is constrained over $\mathcal{E}_\alpha^+$ because the integration over $\mathcal{E}_\alpha$ is dominated by the one over $\mathcal{E}_\alpha^+$. The reason is as follows. Suppose only $\alpha_{n_1}, \dots, \alpha_{n_j} > 0$, $j < m$, and the other α's, $\alpha_{k_1}, \dots, \alpha_{k_{m-j}}$, are negative. Then

$$\prod_{i=1}^{m} P(g_{M-m+i}^2 < \rho^{-\alpha_i}) \approx \prod_{i=1}^{j} P(g_{M-m+n_i}^2 < \rho^{-\alpha_{n_i}}) \approx \rho^{-\sum_{i=1}^j (M-m+n_i)\alpha_{n_i}}.$$

Let $\tilde{\mathcal{E}}_\alpha^+$ denote $\{\{\alpha_{n_i}\}_{i=1}^j > 0 : \sum_{i=1}^j \alpha_{n_i} > m - \sum_{i=1}^{m-j} \alpha_{k_i}\}$. Clearly,

$$\inf_{\tilde{\mathcal{E}}_\alpha^+} \sum_{i=1}^{j} (M-m+n_i)\alpha_{n_i} > \inf_{\mathcal{E}_\alpha^+} \sum_{i=1}^{m} (M-m+i)\alpha_i,$$

which implies that the integration over $\mathcal{E}_\alpha$ is dominated by that over $\mathcal{E}_\alpha^+$. Solving the optimization problem of (4.60) yields

$$d_{\text{GMD}}(M,m) = (M-m+1)m. \tag{4.61}$$

Now we consider UCD. We observe that the power allocation applied to each eigen subchannel is no greater than ρ. Hence the overall channel throughput of UCD is

$$\sum_{i=1}^{m} \log\left(1+\frac{\rho}{m}\lambda_{H,i}^2\right) \le R_{\text{UCD}} \le \sum_{i=1}^{m} \log\left(1+\rho\lambda_{H,i}^2\right), \tag{4.62}$$

where the left term denotes the channel throughput associated with uniform power allocation. Applying UCD, we obtain m subchannels with the same SNR:

$$\sqrt[m]{\prod_{i=1}^{m}\left(1+\frac{\rho}{m}\lambda_{H,i}^2\right)} - 1 \le \rho_{\text{UCD}} \le \sqrt[m]{\prod_{i=1}^{m}\left(1+\rho\lambda_{H,i}^2\right)} - 1. \tag{4.63}$$

The typical error event is

$$\mathcal{E} = \left\{\{\lambda_{H,i}\}_{i=1}^{m} : \rho_{\text{UCD}} < 1\right\}. \tag{4.64}$$

It follows from (4.63) that

$$P_1(\rho) \triangleq P\left(\sqrt[m]{\prod_{i=1}^{m}\left(1+\frac{\rho}{m}\lambda_{H,i}^2\right)} - 1 < 1\right) \ge P(\mathcal{E}) \ge P\left(\sqrt[m]{\prod_{i=1}^{m}\left(1+\rho\lambda_{H,i}^2\right)} - 1 < 1\right) \triangleq P_2(\rho). \tag{4.65}$$

It is easy to see that

$$\lim_{\rho\to\infty} \frac{\log P_1(\rho)}{\log\rho} = \lim_{\rho\to\infty} \frac{\log P_2(\rho)}{\log\rho}. \tag{4.66}$$

Hence

$$\lim_{\rho\to\infty} \frac{\log P(\mathcal{E})}{\log\rho} = \lim_{\rho\to\infty} \frac{\log P_1(\rho)}{\log\rho}, \tag{4.67}$$

which implies that water filling does *not* help improve diversity gain.

It follows from the analyses of [16] that the UCD scheme achieves the optimal diversity-multiplexing tradeoff. In particular, when the transmission data rate is fixed, disregard the increase of input SNR, the diversity gain is $d_{\text{ucd}}(M,m) = Mm$.

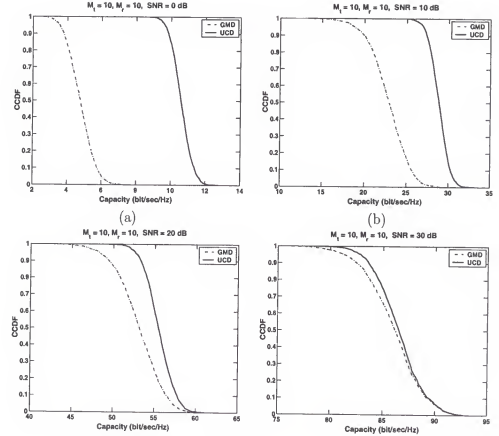

Figure 4–1: Complementary cumulative distribution function of the capacity of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$. Results based on 2000 Monte Carlo trials. SNR = (a) 10 dB, (b) 10 dB (c) 20 dB, and (d) 30 dB.

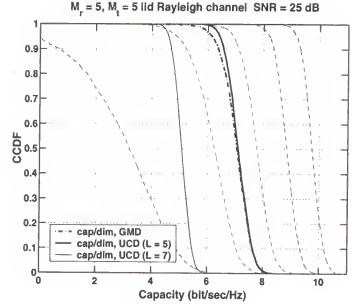

Figure 4–2: Complementary cumulative distribution functions of the capacities of 5 subchannels of an i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 5$. Results based on 2000 Monte Carlo trials.

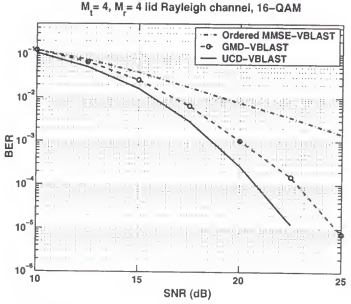

Figure 4–3: Uncoded BER performance when using 16-QAM. Results based on 1000 Monte Carlo trials of an i.i.d. Rayleigh flat fading channel with $M_t = 4$ and $M_r = 4$.

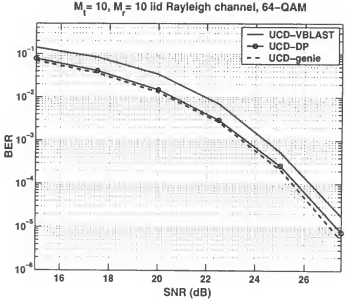

Figure 4–4: BER performances of the UCD-DP, UCD-VBLAST schemes and the imaginary UCD-genie scheme. Results based on 1000 Monte Carlo trials of an i.i.d. Rayleigh flat fading channel with $M_t = 10$ and $M_r = 10$.

CHAPTER 5
TUNABLE CHANNEL DECOMPOSITION

5.1 Introduction

All these aforementioned MIMO transceiver designs focus on improving the communication quality subject to power constraints. In this chapter, we tackle a new aspect of the MIMO transceiver design problem. We regard a MIMO transceiver design as a way of decomposing a MIMO channel into multiple subchannels. As we have mentioned, the MIMO channel decomposition through SVD plus "water filling" lacks flexibility despite its optimality in terms of achieving the maximal overall channel capacity. The success of UCD motivates a much more flexible channel decomposition approach, namely the tunable channel decomposition (TCD) scheme, which is the main result of this chapter. Using the recently developed generalized triangular decomposition (GTD), we propose the TCD scheme to decompose a MIMO channel into multiple subchannels with prescribed capacities or, equivalently, signal-to-interference-and-noise ratios (SINR). The main properties of the TCD scheme are summarized as follows:

1. Given K parallel subchannels with capacities $C_1, C_2, \ldots, C_K$, which is obtained through applying SVD plus "water filling" to a rank K MIMO channel, TCD can convert the K subchannels into $L \geq K$ subchannels with capacities $R_1, R_2, \ldots, R_L$ if and only if $(C_1, \ldots, C_K, 0, \ldots, 0) \in \mathbb{R}_+^L$ majorizes $(R_1, R_2, \ldots, R_L)$ [1] . In particular, $\sum_{i=1}^{K} C_i = \sum_{i=1}^{L} R_i$, i.e., the TCD is capacity lossless.
2. The TCD scheme has two implementation forms. One is the combination of a linear precoder and a minimum mean-squared-error VBLAST (MMSE-VBLAST) detector, which is referred to as TCD-VBLAST, and the other includes a DP precoder and a linear equalizer followed by a DP decoder, which we refer to as TCD-DP.
3. Given the SVD of the MIMO channel matrix, the computational complexity of TCD, which is to calculate the precoder and equalizer matrices, is $O(KL)$, which is computationally quite efficient.

[1] The concept of majorization is introduced in Section 5.2.

Almost originated at the same time as the research on MIMO transceiver designs, the optimal design of symbol synchronous CDMA (S-CDMA) sequences has been under intensive study over the past decade (see, e.g., [31][54][55][56]). Although the two research topics have been studied in an apparently independent manner in the signal processing and information theory communities, the CDMA sequence design problem can be viewed as a special case of the MIMO transceiver design as we have shown in Section 2.1.1. Hence the TCD scheme can be applied, with little modifications, to the design of optimal CDMA sequences. Moreover, the TCD-VBLAST and TCD-DP schemes can be applied to design optimal CDMA sequences in the uplink (mobile-to-base) and downlink (base-to-mobile) scenarios, respectively. Our TCD scheme, which is independently motivated by the MIMO transceiver design problem, turns out to be related to the scheme proposed in [56]. The relationship is discussed in Section 5.3.

5.2 Channel Model and Preliminaries

5.2.1 Channel Model

To facilitate the discussion, we rewrite the channel model used in the previous chapters.

$$\mathbf{y} = \mathbf{H}\mathbf{F}\mathbf{x} + \mathbf{z}, \tag{5.1}$$

where $\mathbf{x} \in \mathbb{C}^{L\times 1}$ is the information symbols precoded by the linear precoder $\mathbf{F} \in \mathbb{C}^{M_t\times L}$ and $\mathbf{y} \in \mathbb{C}^{M_r\times 1}$ is the received signal and $\mathbf{H} \in \mathbb{C}^{M_r\times M_t}$ is the channel matrix with rank K. We assume $E[\mathbf{x}\mathbf{x}^*] = \sigma_x^2\mathbf{I}_L$ and $\mathbf{z} \sim N(0, \sigma_z^2\mathbf{I}_{M_r})$ is the circularly symmetric complex Gaussian noise. We define the SNR as

$$\rho = \frac{E[\mathbf{x}^*\mathbf{F}^*\mathbf{F}\mathbf{x}]}{\sigma_z^2} = \frac{\sigma_x^2}{\sigma_z^2}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\} \triangleq \frac{1}{\alpha}\mathrm{Tr}\{\mathbf{F}^*\mathbf{F}\}, \tag{5.2}$$

5.2.2 Channel Decomposition

Denote the SVD of a rank K channel $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$, where $\mathbf{\Lambda}$ is a $K \times K$ diagonal matrix whose diagonal elements $\{\lambda_{H,k}\}_{k=1}^K$ are the nonzero singular values of $\mathbf{H}$. To maximize the channel capacity with respect to $\mathbf{F}$ given the input power constraint $\mathrm{Tr}\{\mathbf{F}\mathbf{F}^*\} \le \rho\sigma_z^2/\sigma_x^2$, one needs to solve

$$C_{IT} = \max_{\mathrm{Tr}\{\mathbf{F}\mathbf{F}^*\}\le\rho\sigma_z^2/\sigma_x^2} \log_2 |\mathbf{I} + \alpha^{-1}\mathbf{H}\mathbf{F}\mathbf{F}^*\mathbf{H}^*|. \tag{5.3}$$

The optimal linear precoder is (cf. (2.8))

$$\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}. \tag{5.4}$$

Here $L = K$ and Φ is diagonal whose kth ($1 \le k \le K$) diagonal element ϕ_k determines the power loaded to the kth subchannel and is found via "water filling" to be

$$\phi_k(\mu) = \left(\mu - \frac{\alpha}{\lambda_{H,k}^2}\right)^+, \tag{5.5}$$

with μ being chosen such that $\sigma_x^2 \sum_{k=1}^{K} \phi_k(\mu) = \rho\sigma_z^2$ and $(a)^+ = \max\{0, a\}$. In this case, we obtain K subchannels with capacities

$$C_k = \log_2\left(1 + \frac{\phi_k}{\alpha}\lambda_{H,k}^2\right) = \left[\log_2\left(\frac{\mu\lambda_{H,k}^2}{\alpha}\right)\right]^+ \text{ bps/Hz}, \quad k = 1, 2, \ldots, K. \tag{5.6}$$

Due to the usually large dynamic range of singular values $\{\lambda_{H,k}\}_{k=1}^{K}$, the SVD decomposes a MIMO channel into multiple parallel eign-subchannels with different channel capacities. Moreover, since the optimal power loading levels are fixed as given in (5.5), the achievable MIMO channel decomposition is rigidly given in (5.6) and it lacks flexibility.

Another way of decomposing a MIMO channel is to use the VBLAST detector [5]. The VBLAST scheme involves *sequential nulling and cancellation* and it decomposes the MIMO channel into K subchannels (or layers as coined in [5]). By changing the ordering of the signal detection, we can get $K!$ subchannel combinations, each of which is capacity lossless [48].

Theoretically, more combinations of subchannels is possible via time sharing (see [57, Ch. 14.3]). Recall that every DBLAST layer sends its data substream across the K transmitting antennas, or VBLAST layers, in a time sharing manner [2]. For example, for a system with $M_t = 2$, the transmitted data are

$$\begin{array}{lccccc} \text{Vertical Layer-I}: & x_1 & y_2 & x_3 & y_4 & \ldots \\ \text{Vertical Layer-II}: & 0 & x_2 & y_3 & x_4 & \ldots \end{array} \tag{5.7}$$

Let x_i and $y_i, i = 1, 2, \ldots$, denote the symbols transmitted through the DBLAST layers I and II, respectively, at time i. The receiver first estimates x_1 and then estimates x_2 by regarding y_2 as interference. The estimates of x_1, x_2 are decoded jointly, which form the output of the diagonal layer I. After subtracting out the effect of x_1, x_2 from the received data, we can estimate and decode y_2, y_3, which form the diagonal layer II. We remark that DBLAST can be viewed as a combination of VBLAST and the time sharing technique, which decomposes the MIMO channel into multiple *identical* subchannels.

However, time sharing can be difficult to implement in practice. For instance, the major difficulty of DBLAST is the requirement of encoding the diagonal layer with short and efficient error correction codes, which limits its practical implementation despite its superb theoretical performance analyzed in [16].

If CSIT is available, more flexible and practical channel decompositions can be achieved. In Chapter 4, we have proposed the UCD scheme which combines the geometric mean decomposition (GMD) developed in Section 6.2 with either an MMSE-VBLAST detector or a DP precoder to decompose the MIMO channel of (5.1) into $L \geq K$ *identical* subchannels. Hence, the UCD scheme can achieve the theoretical performance of the DBLAST scheme without resorting to any error correcting coding.

In this chapter, we generalize the results of Chapter 4 and develop a systematic channel decomposition that combines the recently proposed GTD algorithm with either an MMSE-VBLAST detector or a DP precoder. We show that given K parallel subchannels with capacities $C_1, C_2, \ldots, C_K$, which are obtained via SVD, TCD can convert the K subchannels into $L \geq K$ subchannels[2] with capacities $R_1, R_2, \ldots, R_L$ if and only if $(R_1, R_2, \ldots, R_L)$ is majorized by $(C_1, \ldots, C_K, 0, \ldots, 0) \in \mathbb{R}^L$. This scheme is particularly relevant to the applications where independent data streams with different qualities-of-service (QoS) share the same MIMO channel [28]. For example, video services usually require higher SNRs than audio services. Decomposing a MIMO channel into multiple subchannels with prescribed capacities and transmitting independent data streams through these subchannels can provide much convenience for resource allocations.

5.2.3 Majorization and Generalized Triangular Decomposition

We introduce several basic concepts and theorems of the majorization theory from [58].

Definition 1 *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, if*

$$\sum_{i=1}^{j} x_{[i]} \leq \sum_{i=1}^{j} y_{[i]}, \quad 1 \leq j \leq n \tag{5.8}$$

with equality holds for $j = n$, where the subscript $[i]$ denotes the ith largest element of the sequence, we say that $\mathbf{x}$ is majorized by $\mathbf{y}$ and denote $\mathbf{x} \prec_+ \mathbf{y}$ or, equivalently, $\mathbf{y} \succ_+ \mathbf{x}$.

[2] If $L < K$, some eign-subchannels are discarded, which causes capacity loss. Hence we focus on the case of $L \geq K$.

Definition 2 *An $n \times n$ matrix $\mathbf{P}$ is doubly stochastic if its (i,j)th entry $p_{ij} \geq 0$ for $i,j = 1,\ldots,n$, and $\sum_{i=1}^{n} p_{ij} = 1$ and $\sum_{j=1}^{n} p_{ij} = 1$.*

Theorem 5.2.1 $\mathbf{x} \prec_+ \mathbf{y}$ *if and only if there exists a doubly stochastic matrix $\mathbf{P}$ such that $\mathbf{x} = \mathbf{P}\mathbf{y}$.*

A square matrix $\mathbf{\Pi}$ is said to be a permutation matrix if each row and column has a single one, and all the other entries are zero. There are $n!$ permutation matrices of size $n \times n$.

Theorem 5.2.2 *The permutation matrices constitute the extreme points of the set of doubly stochastic matrices. Moreover, the set of doubly stochastic matrices is the convex hull of the permutation matrices.*

It follows from Theorems 5.2.1 and 5.2.2 that the set $\{\mathbf{x} \,|\, \mathbf{x} \prec_+ \mathbf{y}\}$ is the convex hull spanned by the $n!$ points which are the permutations of $\mathbf{y}$.

As we have mentioned before, given K parallel subchannels with capacities $C_1, C_2, \ldots, C_K$, which are obtained via SVD, TCD can convert the K subchannels into $L \geq K$ subchannels with capacities $R_1, R_2, \ldots, R_L$ if and only if $(R_1, R_2, \ldots, R_L) \prec_+ (C_1, \ldots, C_K, 0, \ldots, 0) \in \mathbb{R}^L$. For example, for a MIMO channel $\mathbf{H}$ with rank $K = 3$, assume that the capacities of the 3 subchannels obtained via SVD are $C_1 \geq C_2 \geq C_3$. If $L = K$, then TCD can decompose the MIMO channel into 3 subchannels with a rate vector $\mathbf{r} = (R_1, R_2, R_3)$ if and only if $\mathbf{r}$ lies in the convex hull

$$Co\left\{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}, \begin{pmatrix} C_1 \\ C_3 \\ C_2 \end{pmatrix}, \begin{pmatrix} C_2 \\ C_1 \\ C_3 \end{pmatrix}, \begin{pmatrix} C_2 \\ C_3 \\ C_1 \end{pmatrix}, \begin{pmatrix} C_3 \\ C_2 \\ C_1 \end{pmatrix}, \begin{pmatrix} C_3 \\ C_1 \\ C_2 \end{pmatrix}\right\}. \tag{5.9}$$

Here Co stands for the convex hull defined as

$$Co\{S\} = \{\theta_1 x_1 + \ldots + \theta_K x_K | x_i \in S, \theta_i > 0, \theta_1 + \ldots + \theta_K = 1\}. \tag{5.10}$$

In general, the "capacity region" is a convex hull defined by $K!$ vertices in a K-dimensional space. Since the TCD is capacity lossless, i.e., $\sum_{i=1}^{K} C_i = \sum_{i=1}^{K} R_i$, the capacity region falls into a $(K-1)$-dimensional hyperplane. The gray area in Figure 5–1 shows the convex hull of (5.9) with $C_1 = 3$, $C_2 = 2$, and $C_3 = 1$. In this case, the 6 vertices lie in the 2-D plane $\{\mathbf{x} : \sum_{i=1}^{3} x_i = 6\}$. An interesting special case is the UCD scheme [59], which achieves the rate vector corresponding to the center of the convex hull, i.e., $\mathbf{r} = (2, 2, 2)$.

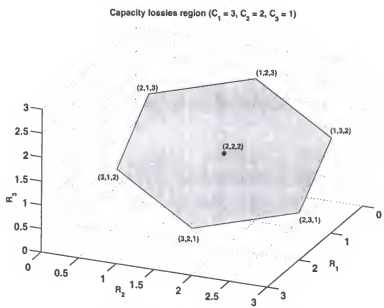

Figure 5–1: Illustration of the capacity lossless region obtainable via TCD. We assume $K = 3$, $C_1 = 3$, $C_2 = 2$, and $C_3 = 1$.

Definition 3 *For* $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^n$, *if*

$$\prod_{i=1}^{j} x_{[i]} \leq \prod_{i=1}^{j} y_{[i]}, \quad 1 \leq j \leq n \tag{5.11}$$

with equality for $j = n$, *we say that* $\mathbf{x}$ *is* multiplicatively *majorized by* $\mathbf{y}$ *and write* $\mathbf{x} \prec_\times \mathbf{y}$ *or, equivalently,* $\mathbf{y} \succ_\times \mathbf{x}$.

Obviously, if $\mathbf{x} \prec_\times \mathbf{y}$, then $\log \mathbf{x} \prec_+ \log \mathbf{y}$.

Now we are ready to introduce the GTD theorem.

Theorem 5.2.3 (GTD theorem) *Let* $\mathbf{H} \in \mathbb{C}^{m \times n}$ *have rank* K *with singular values* $\boldsymbol{\lambda} \in \mathbb{R}_+^K$. *There exists an upper triangular matrix* $\mathbf{R} \in \mathbb{C}^{K \times K}$ *and matrices* $\mathbf{Q}$ *and* $\mathbf{P}$ *with orthonormal columns such that* $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^*$ *if and only if the diagonal elements of* $\mathbf{R}$ *satisfy* $|\mathbf{r}| \prec_\times \boldsymbol{\lambda}$.

Proof: We relegate the proof to Chapter 6. ∎

There is a computationally efficient and numerically stable algorithm to achieve the GTD predicted by Theorem 5.2.3, which is presented in Chapter 6.

5.3 Tunable Channel Decomposition

5.3.1 TCD-VBLAST

We see from (5.2) that $\mathbf{F}$ can always be scaled such that $\alpha = 1$. Hence without loss of generality, we let $\alpha = 1$ in the sequel to simplify the notation.

Denote the SVD of a rank K channel $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^*$, where $\mathbf{\Lambda}$ is a $K \times K$ diagonal matrix whose diagonal elements $\{\lambda_{H,k}\}_{k=1}^K$ are the nonzero singular values of $\mathbf{H}$. The conventional SVD based linear transceiver designs have precoder $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$ where $\mathbf{\Phi}$ is a diagonal matrix whose diagonal elements stand for the power allocation. The precoder $\mathbf{F}$ transforms the MIMO channel into K orthogonal subchannels with capacities

$$C_k = \log_2(1 + \lambda_k^2 \phi_k) \quad \text{bps/Hz}, \quad k = 1, 2, \ldots, K. \tag{5.12}$$

For this kind of precoder design, the only way of controlling the capacity of the subchannels is to change the power allocation $\mathbf{\Phi}$.

If we modify the precoder $\mathbf{F}$ to be [3]

$$\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}} \tag{5.13}$$

where $\mathbf{\Omega} \in \mathbb{R}^{L \times K}$ with $L \geq K$, and $\mathbf{\Omega}^{\mathsf{T}}\mathbf{\Omega} = \mathbf{I}$, then it can been readily seen that introducing $\mathbf{\Omega}$ does not change the overall channel capacity. However, it brings much greater flexibility as demonstrated in the following theorem.

Theorem 5.3.1 (TCD Theorem) *Consider a MIMO channel of (4.1) with $\mathbf{F}$ given in (5.13). For any $L \geq K$, let $\mathbf{c} \in \mathbb{R}^L$ be a zero vector with its first K elements replaced with $\{C_k\}_{k=1}^K$, where $C_k = \log\left(1 + \lambda_{H,k}^2 \phi_k\right)$. Given any rates $\{R_k\}_{k=1}^L$, we can find an orthonormal matrix $\mathbf{\Omega} \in \mathbb{R}^{L \times K}$ such that the combination of the linear precoder $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ and the MMSE-VBLAST detector yields L subchannels with capacities $\{R_k\}_{k=1}^L$ if and only if $\{R_k\}_{k=1}^L \prec_+ \mathbf{c}$.*

Proof: Given the precoder of (5.13), the virtual channel is

$$\mathbf{G} \triangleq \mathbf{H}\mathbf{F} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}} \triangleq \mathbf{U}\mathbf{\Lambda}_G\mathbf{\Omega}^{\mathsf{T}} \tag{5.14}$$

where $\mathbf{\Lambda}_G = \mathbf{\Lambda}\mathbf{\Phi}^{1/2}$ is a diagonal matrix with diagonal elements

$$\lambda_{G,i} = \lambda_{H,i}\phi_i^{1/2}, \quad i = 1, \ldots, K. \tag{5.15}$$

[3] Letting $\mathbf{\Omega}$ to be complex-valued does not introduce additional flexibility as is clear according to the GTD algorithm.

Let the augmented matrix $\mathbf{G}_a$ be defined as

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{U}\Lambda_G\Omega^{\mathsf{T}} \\ \mathbf{I}_L \end{bmatrix}_{(M_r+L)\times L} \tag{5.16}$$

After some straightforward calculations, we can obtain the SVD of $\mathbf{G}_a$ as the following:

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{U}[\Lambda_G \vdots \mathbf{0}_{K\times(L-K)}]\Lambda_{G_a}^{-1} \\ \Omega_0\Lambda_{G_a}^{-1} \end{bmatrix} \Lambda_{G_a}\Omega_0^{\mathsf{T}}, \tag{5.17}$$

where $\mathbf{\Omega}_0 \in \mathbb{R}^{L\times L}$ is orthogonal with its first K columns forming Ω and the diagonal matrix Λ_{G_a} contains the singular values of $\mathbf{G}_a$:

$$\lambda_{G_a,i} = \begin{cases} \sqrt{1+\lambda_{H,i}^2\phi_i}, & 1 \le i \le K, \\ 1, & i > K. \end{cases} \tag{5.18}$$

According to Theorem 5.2.3, we can apply GTD to obtain

$$\Lambda_{G_a} = \mathbf{Q}\mathbf{R}_{G_a}\mathbf{P}^{\mathsf{T}} \tag{5.19}$$

if and only if the diagonal elements of $\mathbf{R}_{G_a} \in \mathbb{R}^{L\times L}$, which we denote as $\{r_{G_a,ii}\}_{i=1}^L$, satisfy

$$\{|r_{G_a,ii}|\}_{i=1}^L \prec_\times \{\lambda_{G_a,i}\}_{i=1}^L. \tag{5.20}$$

Note that both $\mathbf{Q}$ and $\mathbf{P}$ in (5.19) are real-valued matrices because Λ_{G_a} is a real-valued diagonal matrix. Inserting (5.19) into (5.17) yields

$$\mathbf{G}_a = \begin{bmatrix} \mathbf{U}[\Lambda_G \vdots \mathbf{0}_{K\times(L-K)}]\Lambda_{G_a}^{-1} \\ \Omega_0\Lambda_{G_a}^{-1} \end{bmatrix} \mathbf{Q}\mathbf{R}_{G_a}\mathbf{P}^{\mathsf{T}}\Omega_0^{\mathsf{T}}. \tag{5.21}$$

Choose $\mathbf{\Omega}_0 = \mathbf{P}^{\mathsf{T}}$ and define

$$\mathbf{Q}_{G_a} = \begin{bmatrix} \mathbf{U}[\Lambda_G \vdots 0_{K\times(L-K)}]\Lambda_{G_a}^{-1} \\ \Omega_0\Lambda_{G_a}^{-1} \end{bmatrix} \mathbf{Q}. \tag{5.22}$$

Then (5.21) can be rewritten as $\mathbf{G}_a = \mathbf{Q}_{G_a}\mathbf{R}_{G_a}$, which is the QR decomposition of $\mathbf{G}_a$. By Lemma 4.1.2, it follows that for $\alpha = 1$, (5.20) is equivalent to

$$\{1+\rho_i\}_{i=1}^L = \{r_{G_a,ii}^2\}_{i=1}^L \prec_\times \{\lambda_{G_a,i}^2\}_{i=1}^L, \tag{5.23}$$

where ρ_i, $1 \le i \le L$, denotes the output SINR of the ith subchannel, and $\lambda_{G_a,i}$ is given in (5.18). If

$$\{R_i\}_{i=1}^{L} = \{\log(1+\rho_i)\}_{i=1}^{L} \prec_+ \{\log \lambda_{G_a,i}^2\}_{i=1}^{L} = \mathbf{c}, \tag{5.24}$$

then (5.20) and (5.23) hold, which implies the existence of $\mathbf{\Omega}$ (the first K columns of $\mathbf{P}^{\mathsf{T}}$).

Conversely, suppose that there exists a semi-unitary matrix $\mathbf{\Omega}$ such that the linear precoder $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ and the MMSE-VBLAST detector yields L subchannels with capacities $\{R_k\}_{k=1}^{L}$. Let $\mathbf{G}_a = \mathbf{Q}_{G_a}\mathbf{R}_{G_a}$ be the QR decomposition. It follows from Theorem 5.2.3 that (5.20) holds. Hence, by (5.23), we conclude that (5.24) holds. ■

The proof of Theorem 5.3.1 is constructive. Indeed, given the SVD of $\mathbf{H}$ and the power loading level $\mathbf{\Phi}^{1/2}$, we only need to calculate $\mathbf{\Lambda}_G$, $\mathbf{\Lambda}_{G_a}$, and the GTD of $\mathbf{\Lambda}_{G_a}$ given in (5.19). Then we immediately obtain the linear precoder

$$\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}} = \mathbf{V}\left[\mathbf{\Phi}^{1/2} : \mathbf{0}_{K\times(L-K)}\right]\mathbf{P}. \tag{5.25}$$

Let $\mathbf{Q}_{G_a}^{u}$ denote the first M_r rows of $\mathbf{Q}_{G_a}$. Then it follows from (5.22) that

$$\begin{aligned} \mathbf{Q}_{G_a}^{u} &= \mathbf{U}[\mathbf{\Lambda}_G : \mathbf{0}_{K\times(L-K)}]\mathbf{\Lambda}_{G_a}^{-1}\mathbf{Q} \\ &= \mathbf{U}\left[\mathbf{\Gamma} : \mathbf{0}_{K\times(L-K)}\right]\mathbf{Q}, \end{aligned} \tag{5.26}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{K\times K}$ is diagonal with its ith diagonal element being $\gamma_i = \frac{\lambda_{G,i}}{\sqrt{\lambda_{G,i}^2+1}}$. According to Lemma 4.1.1, the nulling vectors are calculated as

$$\mathbf{w}_i = r_{G_a,ii}^{-1}\mathbf{q}_{G_a,i}, \quad 1 \le i \le L, \tag{5.27}$$

where $r_{G_a,ii}$ is the ith diagonal element of $\mathbf{R}_{G_a}$ and $\mathbf{q}_{G_a,i}$ is the ith column of $\mathbf{Q}_{G_a}^{u}$.

In the GTD algorithm, $\mathbf{P}$ and $\mathbf{Q}$ are obtained via multiplication of $L-1$ Givens rotation matrices. Hence calculating (5.25) and (5.26) needs $O(M_t(L+K))$ and $O(M_r(L+K))$ flops, respectively. We note that the decoding starts with the Lth layer, then the $L-1$th, and so on.

Given the SVD of $\mathbf{H}$ and the power allocation level $\mathbf{\Phi}$, the TCD-VBLAST scheme needs to run the procedures summarized in Table 5–1. If $M_t = M_r = K$, then the TCD-VBLAST scheme requires only $O(L^2 + K^2 + KL)$ flops, given the SVD of the channel matrix.

Table 5-1: The TCD-VBLAST Scheme

step	operation	flops
1	Calculate $\mathbf{\Lambda}_G = \mathbf{\Lambda}\mathbf{\Phi}^{1/2}$	$O(K)$
2	Obtain $\mathbf{\Lambda}_{G_a}$ using (5.18)	$O(K)$
3	Apply GTD to $\mathbf{\Lambda}_{G_a}$ to obtain (5.19)	$O(L^2)$
4	Generate $\mathbf{F}$ using (5.25)	$O(M_t(L+K))$
5	Compute $\mathbf{Q}_{G_a}^u$ using (5.26)	$O(M_r(L+K))$
6	Calculate $\{\mathbf{w}_i\}_{i=1}^L$ using (5.27)	$O(M_rL)$

5.3.2 TCD-DP

Similar to UCD, the TCD also have two implementation forms, which are dual to each other. As a dual form of TCD-VBLAST, the TCD scheme can be implemented by using a DP precoder, which we refer to as TCD-DP. For TCD-DP, a direct construction of the linear precoder $\mathbf{F}$ as done in Section 5.3.1 is not obvious. Instead, we exploit the uplink-downlink duality revealed in [49] to obtain TCD-DP. This technique is also used in [59].

We first apply the TCD-VBLAST scheme to the reverse channel

$$\mathbf{y} = \mathbf{H}^*\mathbf{F}\mathbf{x} + \mathbf{z}, \tag{5.28}$$

where the roles of the transmitter and receiver are exchanged and the $\mathbf{H}$ in (5.1) is replaced by $\mathbf{H}^*$. Then we obtain the precoder $\mathbf{F}$ and the equalizer $\mathbf{W} \triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ from $\mathbf{H}^*$ according to (5.25) and (5.27), respectively. Applying $\mathbf{F}$ and the VBLAST detector with nulling vectors $\{\mathbf{w}_i\}_{i=1}^L$, we obtain L subchannels

$$\mathbf{w}_i^*\mathbf{y} = \mathbf{w}_i^*\mathbf{H}^*\mathbf{f}_i x_i + \sum_{j=1}^{i-1} \mathbf{w}_i^*\mathbf{H}^*\mathbf{f}_j x_j + \mathbf{w}_i^*\mathbf{z}, \quad i = 1, \ldots, L, \tag{5.29}$$

where the ith subchannel (5.29) is free of interference from the jth ($j > i$) subchannels which are detected and cancelled out in advance. The SINR of the subchannel (5.29) is

$$\rho_i = \frac{|\mathbf{w}_i^*\mathbf{H}^*\mathbf{f}_i|^2\sigma_x^2}{\mathbf{w}_i^*\mathbf{w}_i\sigma_z^2 + \sum_{j=1}^{i-1}|\mathbf{w}_i^*\mathbf{H}^*\mathbf{f}_j|^2\sigma_x^2}. \tag{5.30}$$

Note that replacing $\mathbf{w}_i$ by $\bar{\mathbf{w}}_i$, which is obtained by scaling $\mathbf{w}_i$ such that $\|\bar{\mathbf{w}}_i\| = 1$, does not change ρ_i since the output SINR is invariant to the length of $\mathbf{w}_i$. Also note that $\alpha = 1$, i.e., $\sigma_z^2 = \sigma_x^2$. Hence (5.29) can be simplified to be

$$\rho_i = \frac{|\bar{\mathbf{w}}_i^*\mathbf{H}^*\mathbf{f}_i|^2}{1 + \sum_{j=1}^{i-1}|\bar{\mathbf{w}}_i^*\mathbf{H}^*\mathbf{f}_j|^2}. \tag{5.31}$$

Let $\bar{\mathbf{f}}_i$, $i = 1, \ldots, L$, be the scaled version of $\mathbf{f}_i$ and has unit length. Denote $p_i = \|\mathbf{f}_i\|^2$. Then

$$\rho_i = \frac{|\bar{\mathbf{w}}_i^* \mathbf{H}^* \bar{\mathbf{f}}_i|^2 p_i}{1 + \sum_{j=1}^{i-1} |\bar{\mathbf{w}}_i^* \mathbf{H}^* \bar{\mathbf{f}}_j|^2 p_j}, \quad i = 1, \ldots, L. \tag{5.32}$$

Let $a_{ij} = |\bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j|^2$. Then (5.32) can be represented in the matrix form

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ -\rho_2 a_{12} & a_{22} & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -\rho_L a_{1L} & -\rho_L a_{2L} & \cdots & a_{LL} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_L \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_L \end{bmatrix}. \tag{5.33}$$

According to the uplink-downlink duality, in the original channel, the precoder of TCD-DP should be $\tilde{\mathbf{F}} = [\sqrt{q_1}\bar{\mathbf{w}}_1, \ldots, \sqrt{q_L}\bar{\mathbf{w}}_L]$, where $\{q_i\}_{i=1}^L$ will be determined later in (5.37), and the receiving vectors are $\bar{\mathbf{f}}_i$, $i = 1, \ldots, L$. Then we get L subchannels whose ith scalar subchannel of the MIMO channel is

$$y_i = \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_i \sqrt{q_i} x_i + \sum_{j=i+1}^{L} \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \sum_{j=1}^{i-1} \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \bar{\mathbf{f}}_i^* \mathbf{z}. \tag{5.34}$$

Applying the dirty paper precoder to x_i and treating $\sum_{j=1}^{i-1} \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j$ as the interference known at the transmitter (note that here we precode the first layer first while for TCD-VBLAST, we detect the Lth layer first), we obtain an equivalent subchannel

$$y_i = \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_i \sqrt{q_i} x_i + \sum_{j=i+1}^{L} \bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j \sqrt{q_j} x_j + \bar{\mathbf{f}}_i^* \mathbf{z} \tag{5.35}$$

with SINR (again, recall that $\alpha = 1$ and $\sigma_x^2 = \sigma_z^2$)

$$\rho_i = \frac{q_i |\bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_i|^2}{1 + \sum_{j=i+1}^{L} q_j |\bar{\mathbf{f}}_i^* \mathbf{H} \bar{\mathbf{w}}_j|^2}, \quad \text{for } i = 1, 2, \ldots, L. \tag{5.36}$$

Similar to (5.32), (5.36) can also be represented as

$$\begin{bmatrix} a_{11} & -\rho_1 a_{12} & \cdots & -\rho_1 a_{1L} \\ 0 & a_{22} & \cdots & -\rho_2 a_{2L} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{LL} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_L \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_L \end{bmatrix}. \tag{5.37}$$

It is easy to see that $q_i > 0$, $0 \le i \le L$. It is proven in [49] that $\sum_{i=1}^{L} q_i = \mathrm{tr}(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^*) = \mathrm{tr}(\mathbf{F}\mathbf{F}^*) = \sum_{i=1}^{L} p_i$. That is, to obtain L subchannels with SINRs $\{\rho_i\}_{i=1}^L$, the TCD-DP

needs exactly the same power as the TCD-VBLAST. To make this chapter self-contained, we give below an alternative proof to this interesting and useful fact.

Let $\mathcal{U}_A$ denote a strictly upper triangular matrix whose (i,j)th entry is a_{ij} for $1 \le i < j \le L$ and zero otherwise. Let $\mathcal{D}_A$ and $\mathcal{D}_\rho$ be two $L \times L$ diagonal matrices with their ith element equal to a_{ii} and ρ_i, respectively. Then (5.32) can be rewritten as

$$\left(\mathcal{D}_A - \mathcal{D}_\rho \mathcal{U}_A^{\mathsf{T}}\right) \mathbf{p} = \boldsymbol{\rho} \tag{5.38}$$

or equivalently

$$\left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A^{\mathsf{T}}\right) \mathbf{p} = \mathbf{1} \tag{5.39}$$

where $\mathbf{p} = [p_1, \ldots, p_L]^{\mathsf{T}}$, $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_L]^{\mathsf{T}}$ and $\mathbf{1}$ is a vector with unit elements. Hence

$$\mathbf{p} = \left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A^{\mathsf{T}}\right)^{-1} \mathbf{1} \tag{5.40}$$

Similarly, (5.37) can be rewritten as

$$\left(\mathcal{D}_A - \mathcal{D}_\rho \mathcal{U}_A\right) \mathbf{q} = \boldsymbol{\rho} \tag{5.41}$$

or

$$\left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A\right) \mathbf{q} = \mathbf{1}. \tag{5.42}$$

Hence

$$\mathbf{q} = \left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A\right)^{-1} \mathbf{1}. \tag{5.43}$$

From (5.40) and (5.43),

$$\sum_{i=1}^{L} p_i = \mathbf{1}^{\mathsf{T}} \left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A^{\mathsf{T}}\right)^{-1} \mathbf{1} = \mathbf{1}^{\mathsf{T}} \left(\mathcal{D}_\rho^{-1}\mathcal{D}_A - \mathcal{U}_A\right)^{-1} \mathbf{1} = \sum_{i=1}^{L} q_i. \tag{5.44}$$

We can use the Tomlinson-Harashima precoder [42][43] or the trellis precoder [44] to achieve known interference cancellation at the transmitter. For a system with high dimensionality, TCD-DP is a better choice than TCD-VBLAST since it is free of propagation errors.

5.4 MIMO Communications with QoS Constraints

In this section, we apply the TCD scheme to MIMO communications with QoS constraints. Suppose we want to transmit $L \ge K$ independent data streams through a MIMO channel. Instead of multiplexing all the substreams in the time division manner to share the entire MIMO channel, we apply TCD to decompose the MIMO channel into multiple subchannels whose capacities/SINRs meet the QoS requirements of the

substreams, and dedicate one subchannel to each substream. In [28], the authors studied the same problem. They proposed a linear transceiver design which, similar to TCD, can also control the SINR of each subchannel via designing the precoder. However, the linear transceiver is capacity lossy and can suffer from considerable performance degradation compared with our TCD scheme as we will show at the end of this section. Given that all the subchannels meet the QoS constraints, we want to minimize the overall input power. We need to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{F}} \quad & \text{tr}\,(\mathbf{F}\mathbf{F}^*) \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{H}\mathbf{F} \\ \mathbf{I}_L \end{pmatrix} = \mathbf{Q}\mathbf{R} \\ & \text{diag}(\mathbf{R}) = \{\sqrt{1+\rho_i}\}_{i=1}^{L}. \end{aligned} \tag{5.45}$$

Here $\mathbf{QR}$ denotes the QR decomposition and $\text{diag}(\mathbf{R})$ denotes the vector formed by the diagonal of $\mathbf{R}$. According to Lemma 4.1.2, the diagonal of $\mathbf{R}$ determines the SINRs of the subchannels. Without loss of generality, we assume that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_L$. We now consider a problem whose constraints are more relaxed than those of (5.45):

$$\begin{aligned} \min_{\mathbf{F}} \quad & \text{tr}\,(\mathbf{F}\mathbf{F}^*) \\ \text{subject to} \quad & \boldsymbol{\lambda}_{G_a} \succ_\times \{\sqrt{1+\rho_i}\}_{i=1}^{L}, \quad \mathbf{G}_a = \begin{pmatrix} \mathbf{H}\mathbf{F} \\ \mathbf{I}_L \end{pmatrix}, \end{aligned} \tag{5.46}$$

where $\boldsymbol{\lambda}_{G_a}$ stands for the singular values of the augmented matrix $\mathbf{G}_a$. In general, for any matrix $\mathbf{A}$, we let $\boldsymbol{\lambda}_A$ denote the singular values of $\mathbf{A}$. By Theorem 5.2.3, if $\mathbf{F}$ is feasible in (5.45), then $\mathbf{F}$ is feasible in (5.46). We now further simplify (5.46) and show that its solution provides a solution of (5.45).

Theorem 5.4.1 *If $\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^*$ is the singular value decomposition of $\mathbf{H}$, then* (5.46) *has a solution of the form $\mathbf{F} = \mathbf{V}\boldsymbol{\Phi}^{1/2}$ where $\boldsymbol{\Phi} \in \mathbb{R}^{K\times K}$ is a diagonal matrix with diagonal elements ϕ_i, $1 \leq i \leq K$, chosen to solve the problem*

$$\left.\begin{aligned} \min_{\boldsymbol{\Phi}} \quad & \textstyle\sum_{i=1}^{K} \phi_i \\ \textit{subject to} \quad & \textstyle\prod_{i=1}^{k}(1+\lambda_{H,i}^2\phi_i) \geq \prod_{i=1}^{k}(1+\rho_i), \quad \phi_k \geq \phi_{k+1} \geq 0, \quad 1 \leq k \leq K-1, \\ & \textstyle\prod_{i=1}^{K}(1+\lambda_{H,i}^2\phi_i) = \prod_{i=1}^{L}(1+\rho_i). \end{aligned}\right\} \tag{5.47}$$

Moreover, if $\mathbf{Q}\mathbf{R}_{G_a}\mathbf{P}^{\mathsf{T}}$ *is the GTD of* $\mathbf{\Lambda}_{G_a}$ *in (5.19), then* (5.45) *has the solution* $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ *where* $\mathbf{\Phi}$ *is a solution of* (5.47) *and* $\mathbf{\Omega}$ *is the matrix formed by the first* K *columns of* $\mathbf{P}^{\mathsf{T}}$.

Proof: See Appendix A. ∎

We now develop an efficient algorithm for solving (5.47). We will see that the constraint $\phi_k \geq \phi_{k+1}$ can be omitted since it is automatically satisfied at a minimizer of (5.47). To begin, we make a change of variables to further simplify the formulation of (5.47). We define

$$\begin{aligned} \psi_i &= \phi_i + 1/\lambda_{H,i}^2, && 1 \leq i \leq K, \\ \beta_i &= \tfrac{1+\rho_i}{\lambda_{H,i}^2}, && 1 \leq i < K, \\ \beta_K &= \tfrac{1}{\lambda_{H,K}^2} \textstyle\prod_{i=K}^{L}(1+\rho_i). \end{aligned} \tag{5.48}$$

With these definitions, (5.47) reduces to

$$\left.\begin{array}{ll} \min_{\boldsymbol{\psi}} & \sum_{i=1}^{K} \psi_i \\ \text{subject to} & \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i, \quad \psi_k \geq 1/\lambda_{H,k}^2, \ 1 \leq k \leq K. \end{array}\right\} \tag{5.49}$$

Both the equality constraint and the inequalities $\phi_k \geq \phi_{k+1}$ in (5.47) have been dropped since these constraints are automatically satisfied at an optimum. The fact that $\phi_k \geq \phi_{k+1}$ is established after Lemma 5.4.2. With regard to the equality constraint, if ψ is feasible in (5.49) and the inequality corresponding to $k = K$ is strictly positive, then the cost is reduced when the trailing components of ϕ are lowered. That is, if $\boldsymbol{\psi}$ is feasible in (5.49) and the inequality corresponding to $k = K$ is strictly positive, then the cost is reduced when the trailing components of $\boldsymbol{\psi}$ are lowered.

Clearly, the feasible set for (5.49) is nonempty and the cost function tends to infinity as any of the components of $\boldsymbol{\psi}$ tends to infinity. By continuity of the cost function and the constraints, a minimizer must exist. We now analyze the structure of the minimizer. By exploiting the structure, we obtain a fast algorithm for solving (5.49).

We first study a similar optimization problem with relaxed constraints.

Lemma 5.4.1 *Any solution* $\boldsymbol{\psi}$ *of the problem*

$$\min_{\boldsymbol{\psi}} \sum_{i=1}^{K} \psi_i \ \ \textit{subject to} \ \ \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i, \ \ 1 \leq k \leq K, \tag{5.50}$$

has the property that $\psi_{i+1} \leq \psi_i$ *for each* i.

Proof: We replace the inequalities in (5.50) by the equivalent constraints obtained by taking log's:

$$\sum_{i=1}^{k} \log(\psi_i) \geq \sum_{i=1}^{k} \log(\beta_i), \quad 1 \leq k \leq K,$$

The Lagrangian $\mathcal{L}$ associated with (5.50), after this modification of the constraints, is

$$\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \left(\psi_k - \mu_k \sum_{i=1}^{k} (\log(\psi_i) - \log(\beta_i)) \right).$$

By the first-order optimality conditions associated with $\boldsymbol{\psi}$, there exists $\boldsymbol{\mu} \geq \mathbf{0}$ with the property that the gradient of the Lagrangian with respect to $\boldsymbol{\psi}$ vanishes. Equating to zero the partial derivative of the Lagrangian with respect to ψ_j, we obtain the relations

$$\psi_j = \sum_{i=j}^{K} \mu_i, \quad j = 1, \ldots, K.$$

Hence, $\psi_j - \psi_{j+1} = \mu_j \geq 0$. ∎

Using Lemma 5.4.1, we can gain insights into the structure of a solution to (5.49).

Lemma 5.4.2 *There exists a solution $\boldsymbol{\psi}$ to* (5.49) *with the property that for some integer $j \in [1, K]$,*

$$\psi_{i+1} \leq \psi_i \text{ for all } i < j, \quad \psi_{i+1} \geq \psi_i \text{ for all } i \geq j, \quad \psi_i = \frac{1}{\lambda_{H,i}^2} \text{ for all } i > j. \tag{5.51}$$

In particular, $\psi_j \leq \psi_i$ for all i.

Proof: If $\boldsymbol{\psi}$ is a solution of (5.49) with the property that $\psi_i > \frac{1}{\lambda_{H,i}^2}$ for all $1 \leq j \leq K$, then by the convexity of the constraints, it follows that $\boldsymbol{\psi}$ is a solution of (5.50). By Lemma 5.4.1, we conclude that Lemma 5.4.2 holds with $j = K$. Now, suppose that $\boldsymbol{\psi}$ is a solution of (5.49) with $\psi_i = 1/\lambda_{H,i}^2$ for some i. We wish to show that $\psi_k = 1/\lambda_{H,k}^2$ for all $k > i$. Suppose, to the contrary, that there exists an index $k \geq i$ with the property that $\psi_k = 1/\lambda_{H,k}^2$ and $\psi_{k+1} > 1/\lambda_{H,k+1}^2$. We show that components k and $k+1$ of $\boldsymbol{\psi}$ can be modified so as to satisfy the constraints and make the cost strictly smaller. In particular, let $\boldsymbol{\psi}(\epsilon)$ be identical with $\boldsymbol{\psi}$ except for components k and $k+1$:

$$\psi_k(\epsilon) = (1+\epsilon)\psi_k \quad \text{and} \quad \psi_{k+1}(\epsilon) = \frac{\psi_{k+1}}{1+\epsilon}. \tag{5.52}$$

For $\epsilon > 0$ small, $\boldsymbol{\psi}(\epsilon)$ satisfies the constraints of (5.49). The change $\Delta(\epsilon)$ in the cost function of (5.49) is

$$\Delta(\epsilon) = (1+\epsilon)\psi_k + \frac{\psi_{k+1}}{1+\epsilon} - \psi_k - \psi_{k+1}.$$

The derivative of $\Delta(\epsilon)$ evaluated at zero is

$$\Delta'(0) = \psi_k - \psi_{k+1}.$$

Since $1/\lambda_{H,k}^2$ is an increasing function of k and since $\psi_k = 1/\lambda_{H,k}^2$, we conclude that $\psi_{k+1} > \psi_k$ and $\Delta'(0) < 0$. Hence, for $\epsilon > 0$ near zero, $\boldsymbol{\psi}(\epsilon)$ has a smaller cost than $\boldsymbol{\psi}(0)$, which yields a contradiction. Hence, there exists an index j with the property that $\psi_i = 1/\lambda_{H,i}^2$ for all $i > j$ and $\psi_i > 1/\lambda_{H,i}^2$ for all $i \leq j$.

According to Lemma 5.4.1, $\psi_i \geq \psi_{i+1}$ for any $i < j$. To complete the proof, we need to show that $\psi_j \leq \psi_{j+1}$. As noted previously, any solution of (5.49) satisfies

$$\prod_{i=1}^{K} \psi_i = \prod_{i=1}^{K} \beta_i,$$

which implies (cf. (5.48))

$$\prod_{i=1}^{j} \psi_i = \prod_{i=1}^{j} \beta_i \left(\prod_{i=j+1}^{K} \beta_i \lambda_{H,i}^2 \right) > \prod_{i=1}^{j} \beta_i.$$

That is, the constraint $\prod_{i=1}^{j} \psi_i \geq \prod_{i=1}^{j} \beta_i$ in (5.49) is inactive. If $\psi_j > \psi_{j+1}$, we will decrease the j-th component and increase the $j+1$ component, while leaving the other components unchanged. Letting $\boldsymbol{\psi}(\delta)$ be the modified vector, we set

$$\psi_{j+1}(\delta) = (1+\delta)\psi_{j+1} \quad \text{and} \quad \psi_j(\delta) = \frac{\psi_j}{1+\delta}.$$

Since the j-th constraint in (5.49) is inactive, $\boldsymbol{\psi}(\delta)$ is feasible for δ near zero. And if $\psi_j > \psi_{j+1}$, then the cost decreases as δ increases. It follows that $\psi_j \leq \psi_{j+1}$. ■

By Lemma 5.4.2, ϕ_i is a decreasing function of i for $i \in [1, j]$ while $\phi_i = 1/\lambda_{H,i}^2$ for $i > j$. Since $\lambda_{H,i}$ is a decreasing function of i, it follows that $\phi_i = \psi_i - 1/\lambda_{H,i}^2$ is a decreasing function of i for $i \in [1, j]$ with $\phi_i \geq 0$, while $\phi_i = 0$ for $i > j$. Hence, ϕ_i is a decreasing function of $i \in [1, K]$. In particular, the constraint $\phi_k \geq \phi_{k+1}$ in (5.47) is automatically satisfied by the associated solution characterized in Lemma 5.4.2.

We refer to the index j in Lemma 5.4.2 as the "break point." At the break point, the lower bound constraint $\psi_i \geq 1/\lambda_{H,i}^2$ changes from inactive to active. We now use Lemma 5.4.2 to obtain an algorithm for (5.49).

Lemma 5.4.3 *Let γ_k denote the k-th geometric mean of the β_i:*

$$\gamma_k = \left(\prod_{i=1}^{k} \beta_i \right)^{1/k},$$

```
function ψ = TCDPow (β,λ)
L = 1 ; R = length (β) ; ψ = zeros (1, R) ;
ζ = cumsum (log (β)) ;
while R ≥ L
     [t, l] = max (ζ(L:R)./[1:R-L+1]) ;
     γl = exp (t) ; L1 = L + l - 1 ;
     if γl ≥ 1/λ(L1)^2
          ψ(L:L1) = γl ;
          L = L + l ;
          ζ(L:R) = ζ(L:R) - ζ(L-1) ;
     else
          ψ(L1:R) = 1./(λ(L1:R).^2) ;
          ζ(L1-1) = ζ(R) - sum (log (ψ(L1:R))) ;
          R = L1 - 1 ;
     end
end
```

Figure 5–2: A Matlab function to solve (5.49).

and let l denote an index for which γ_k is the largest:

$$l = \arg\max\{\gamma_k : 1 \le k \le K\}. \tag{5.53}$$

If $\gamma_l \ge 1/\lambda_l^2$, then putting $\psi_i = \gamma_l$ for all $i \le l$ is optimal in (5.49). *If $\gamma_l < 1/\lambda_l^2$, then $\psi_i = 1/\lambda_{H,i}^2$ for all $i \ge l$ at an optimal solution of* (5.49).

Proof: See Appendix B. ∎

Based on Lemma 5.4.3, we can use the following strategy to solve (5.49). We form the geometric mean described in Lemma 5.4.3 and we evaluate l. If $\gamma_l \ge 1/\lambda_{H,l}^2$, then we set $\phi_i = \gamma_l$ for $i \le l$, and we simplify (5.49) by removing ψ_i, $1 \le i \le l$, from the problem. If $\gamma_l < 1/\lambda_{H,l}^2$, then we set $\psi_i = 1/\lambda_{H,i}^2$ for $i \ge l$, and we simplify (5.49) by removing ψ_i, $l \le i \le K$, from the problem. The Matlab code `TCDPow` implementing this algorithm appears in Figure 5–2.

After obtaining the power loading level $\phi_i = \psi_i - 1/\lambda_{H,i}^2$, $1 \le i \le K$, we calculate the precoder $\mathbf{F}$ and the nulling vectors $\{\mathbf{w}_i\}_{i=1}^L$ according to Table 5–1 in Section 5.3. Note that one of the possible paths through the `TCDPow` routine makes the leading elements of $\boldsymbol{\psi}$ all equal while setting the trailing elements of $\psi_i = 1/\lambda_{H,i}^2$. This path coincides with the standard water filling algorithm. In this case, the TCD scheme is optimal in terms of maximizing the overall throughput given the input power. On the other hand, if some substream has a very high prescribed SINR such that the l given in (5.53) is less than the "break point" j, then $\boldsymbol{\psi}$ leads to be a multi-level water filling power allocation, which

suffers from overall capacity loss. This happens when the target rate vector $[R_1, \ldots R_L]$ falls out of the convex hull spanned by the $L!$ permutations of $[C_1, \ldots, C_K, 0, \ldots, 0]$ (cf. Figure 5–1), where $C_k, k = 1, \ldots, K$, are the capacities of the eigen subchannels with water filling power allocation. As a remedy to this issue, one can "break" (if it is practically allowable) the *oversized* substream into more than one substreams with smaller rates, or equivalently, lower SINR requirements. Note that TCD can decompose a MIMO channel into an arbitrarily large number of subchannels.

An interesting special case is that $\rho_1 = \rho_2 = \ldots = \rho_L$, i.e., the substream shares the same SINR requirements. In this case, $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_K$ since the singular values $\{\lambda_{H,i}\}_{i=1}^{K}$ are in nonincreasing order, and `TCDPow` yields a standard water filling solution. In this case, TCD becomes UCD.

We present two numerical example to conclude this section. In the first example, we assume Rayleigh independent flat fading channels with $M_t = 5$ and $M_r = 6$. We consider equal QoS requirements for $L = 5$ independent substreams. Figure 5–3 compares the input power needed by our TCD scheme and the linear transceiver scheme of [28]. Our scheme can save about 2.5 dB for any prescribed output SINR.

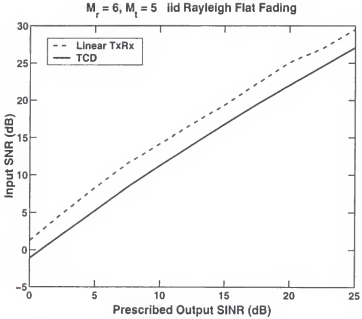

Figure 5–3: Input SNR vs. Output SINR. The result is based on the average of 500 Monte Carlo trials of a i.i.d. Rayleigh flat fading channel with $M_t = 5$ and $M_r = 6$.

In the second example, we consider a rank two MIMO channel with singular value λ_1, λ_2. Suppose we want to decompose the MIMO channel into 2 subchannels with

capacity C_1 and C_2 with $C_1 + C_2 = 10$ bps/Hz. We consider the three scenarios with $(\lambda_1 = 2, \lambda_2 = 1)$, $(\lambda_1 = 5, \lambda_2 = 1)$, and $(\lambda_1 = 10, \lambda_2 = 1)$. For all the three cases, there is an inflection point beyond which our TCD is the same as the linear design of [28]. That is because when the two subchannels have very disparate QoS constraints, i.e., C_1 is far larger than C_2, the optimal strategy is to apply SVD to the channel matrix and transmit data through the orthogonal eign-subchannels. (In this case, $\mathbf{\Omega} = \mathbf{I}$. (cf. (5.13)).) If the subchannels QoS constraints are not too disparate, which corresponds to the region to the left of the inflection point, the required input power of our TCD scheme is invariant with respect to C_1, C_2 and is strictly less than that needed by the linear design. This region corresponds to the capacity lossless region (cf. Figure 5–1). Another interesting point is that the relative advantage of TCD is more prominent if the singular values λ_1, λ_2 become more disparate.

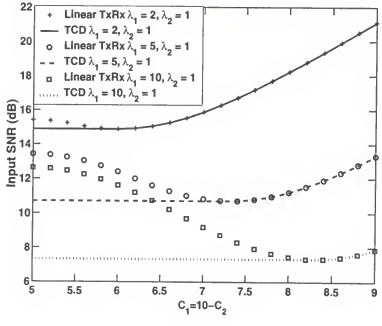

Figure 5–4: Input SNR vs. C_1. A rank 2 channel is decomposed into two subchannels with capacities C_1 and $C_2 = 10 - C_1$.

5.5 CDMA Sequence Design

As we have shown in Section 2.1.1, the CDMA sequence design problem can be viewed as a special case of the MIMO transceiver design. In an idealized S-CDMA system where the channel does not experience any fading or near-far effect, L mobile users modulate their information symbols via spreading sequences $\{\mathbf{s}_i\}_{i=1}^{L}$, each of which has the processing gain N. The discrete-time baseband S-CDMA signal received at the

(single-antenna) base-station can be represented as [31]

$$\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{z} \tag{5.54}$$

where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_L] \in \mathbb{R}^{N \times L}$ and the lth ($1 \leq l \leq L$) entry of $\mathbf{x}$, x_l, stands for the information symbol from the lth user. In the downlink channel, the base station multiplexes the information dedicated to the L mobile users through the spreading sequences, which are the columns of $\mathbf{S}$. Then, all the mobiles receive the same signal given in (5.54). We remark that (5.54) can also be written as (4.1) with $\mathbf{H} = \mathbf{I}_N$ and $\mathbf{F} = \mathbf{S}$. Here $M_r = M_t = N$ is the processing gain. Hence, optimizing the spreading sequences amounts to optimizing the precoder $\mathbf{F}$ for a MIMO system. Indeed, due to the simple channel matrix ($\mathbf{H} = \mathbf{I}$), some procedures of the TCD scheme can be simplified. We shall show that the TCD scheme turns out to be an improved solution to the sequence design proposed in [56]. At the end of this section, we will compare our TCD scheme and the scheme proposed in [56].

5.5.1 CDMA Sequences Maximizing Sum Capacity

Recall that the precoder maximizing the overall MIMO channel capacity is $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ where $\mathbf{\Phi}$ is obtained by water filling algorithm. For an S-CDMA channel, $\mathbf{H} = \mathbf{I}$, then $\mathbf{V} = \mathbf{I}$ and the optimal power loading level is the uniform power allocation. Hence the CDMA sequence maximizing the sum capacity is $\mathbf{S} = \sqrt{\rho}\mathbf{\Omega}^{\mathsf{T}}$. Since $\mathbf{\Omega}$ has orthonormal columns, we obtain $\mathbf{S}\mathbf{S}^{\mathsf{T}} = \rho\mathbf{I}$. This observation coincides with the findings in [31], in which the authors show that the CDMA sequences maximizing the sum capacity are the Welch-Bound-Equality sequences.

5.5.2 Uplink Case

For the uplink scenario, i.e., the mobiles to base station case, the base station calculates the optimal CDMA sequences for each mobile user and the associated successive nulling vectors needed by itself. Then the base station informs the mobile users their designated CDMA sequences.

First, we need to calculate the power loading levels $\mathbf{\Phi} \in \mathbb{R}^{N \times N}$ such that the following GTD matrix decomposition is possible:

$$\mathbf{\Phi}_a \triangleq \begin{pmatrix} [\mathbf{\Phi}^{1/2} : \mathbf{0}_{N \times (L-N)}] \\ \mathbf{I}_L \end{pmatrix} = \mathbf{Q}\mathbf{R}\mathbf{P}^{\mathsf{T}}, \tag{5.55}$$

where the diagonal elements of $\mathbf{R}$, $r_{ii}, i = 1, 2, \ldots, L$, satisfy the QoS constraints. Note that the singular values of $\mathbf{\Phi}_a$ form a sequence whose first N elements are $\sqrt{1+\phi_i}, i =$

$1, 2, \ldots, N$, followed by $L - N$ ones. From Theorem 5.2.3, (5.55) exists if and only if

$$(\{1+\phi_i\}_{i=1}^{N}, 1, \ldots, 1) \succ_{\times} \{1+\rho_i\}_{i=1}^{L}. \tag{5.56}$$

Similar to (5.47), we need to solve the problem

$$\begin{array}{ll} \min_{\Phi} & \sum_{i=1}^{N} \phi_i \\ \text{subject to} & (\{1+\phi_i\}_{i=1}^{N}, 1, \ldots, 1) \succ_{\times} \{1+\rho_i\}_{i=1}^{L}. \\ & \phi_i \geq 0, \quad \forall i \end{array} \tag{5.57}$$

Similar to (5.49), (5.57) can be further simplified using the variables

$$\psi_i = \phi_i + 1, \quad \beta_i = 1 + \rho_i \text{ for } i < N, \quad \text{and} \quad \beta_N = \prod_{i=N}^{L} (1+\rho_i).$$

The simplified problem is

$$\left.\begin{array}{ll} \min_{\Psi} & \sum_{i=1}^{N} \psi_i \\ \text{subject to} & \prod_{i=1}^{k} \psi_i \geq \prod_{i=1}^{k} \beta_i, \quad \psi_k \geq 1, \ \ 1 \leq k \leq N. \end{array}\right\} \tag{5.58}$$

The algorithm `TCDPow` simplifies immensely when we apply it to (5.58). Since $\beta_i \geq 1 = \frac{1}{\lambda_{H,i}}$ for all i, the constraints $\psi_i \geq 1$ are inactive. Since $\beta_i \leq \beta_{i-1}$ for all $i < N$, the geometric means satisfy $\gamma_i \leq \gamma_{i-1}$ for all $i < N$. Hence, in Lemma 5.4.3, the value of l is either 1 or N. If $l = 1$, then we set $\psi_1 = \beta_1$ and we remove ψ_1 from the problem. If $l = N$, then $\psi_i = \gamma_N$ for all i. It follows that there exists an index j with the property that

$$\psi_i = \beta_i \text{ for all } i \leq j \quad \text{and} \quad \psi_i = \left(\prod_{i=j+1}^{N} \beta_i \right)^{1/(N-j)} \quad \text{for all } i > j.$$

This observation coincides with the solution obtained in [56].

Let $\boldsymbol{\Psi}$ denote an $L \times L$ identity matrix with its first N diagonal elements replaced by $\psi_i, 1 \leq i \leq N$. According to the TCD scheme presented in Section 5.3.1, we then apply the GTD algorithm to $\boldsymbol{\Psi}^{1/2}$ to obtain

$$\boldsymbol{\Psi}^{1/2} = \mathbf{Q}_{\Psi} \mathbf{R}_{\Psi} \mathbf{P}_{\Psi}^{\mathrm{T}}. \tag{5.59}$$

According to (5.25),

$$\mathbf{S} = \mathbf{F} = \left[\boldsymbol{\Phi}^{1/2} \vdots \mathbf{0}_{N \times (L-N)} \right] \mathbf{P}_{\Psi}. \tag{5.60}$$

Let

$$[\mathbf{v}_1, \ldots, \mathbf{v}_L] = [\boldsymbol{\Phi}^{1/2} \vdots \mathbf{0}_{N \times (L-N)}] \boldsymbol{\Psi}^{-1/2} \mathbf{Q}_{\Psi}. \tag{5.61}$$

By (5.26) and (5.27), the nulling vectors used at the base station are

$$\mathbf{w}_i = r_{\Psi,ii}^{-1}\mathbf{v}_i,\ i = 1, 2, \ldots, L, \tag{5.62}$$

where $r_{\Psi,ii}$ is the ith diagonal element of $\mathbf{R}_\Psi$. In summary, the base station needs to run the following three steps:

1. Solve the optimization problem (5.58).
2. Apply the GTD algorithm to $\Psi^{1/2}$ in (5.59).
3. Obtain the spreading sequences for all mobile users, $[\mathbf{s}_1, \ldots, \mathbf{s}_L] = \mathbf{S}$, and the nulling vectors $\{\mathbf{w}_i\}_{i=1}^L$ (cf. (5.60) and (5.62)) for the base station.

5.5.3 Downlink Case

In the downlink case, the mobiles cannot cooperate with each other for decision feed-back. Hence the VBLAST detection is impractical at receivers. However, we can apply TCD-DP as introduced in Section 5.3.2 to cancel out known interferences at the transmitter, i.e., the base station. We can convert the downlink problem as an uplink one and exploit the downlink-uplink duality as we have done in Section 5.3.2. Note that $\mathbf{H} = \mathbf{H}^* = \mathbf{I}$, i.e., the downlink and uplink channels are the same! Consider the case where the uplink and downlink communications are symmetric, i.e., for each mobile user, the QoS of the communications from the user to the base station and the base station to the user are the same. After obtaining the spreading sequences $[\mathbf{s}_1, \ldots, \mathbf{s}_L]$ for the mobile users, and the nulling vectors $[\mathbf{w}_1, \ldots, \mathbf{w}_L]$ used at the base station for the uplink case, we immediately know that in the downlink case the spreading sequences transmitted from the base station are exactly $[\mathbf{w}_1, \ldots, \mathbf{w}_L]$ and the nulling vectors used at the mobiles are the spreading sequences, $[\mathbf{s}_1, \ldots, \mathbf{s}_L]$, used in the uplink case. The only parameters we need to calculate are $q_1, \ldots, q_N$ (cf. (5.37)). Hence in this symmetric case, the base station only needs to inform the mobiles their designated spreading sequences once in the two-way communications. Each mobile uses the same sequence for both data transmission in the uplink channel and interference nulling in the downlink channel.

5.5.4 Numerical Example

We present one numerical example to show how TCD can be applied to CDMA sequence design. We consider an example where there are $L = 4$ mobile users and the processing gain $N = 3$. The prescribed SINRs of the four users are 20, 19, 18, and 17 dB, respectively. For the uplink case, we apply the TCD-VBLAST scheme to obtain

the spreading sequences of the four users as the columns of the matrix

$$\mathbf{S} = \begin{pmatrix} 10.0000 & -12.0745 & -6.4974 & -3.0926 \\ 0 & 0 & 7.4138 & -15.5760 \\ 0 & 8.8312 & -13.3801 & -6.3686 \end{pmatrix}. \tag{5.63}$$

The nulling vectors used by the base station are the columns of the matrix

$$\mathbf{W} = \begin{pmatrix} 0.0990 & -0.0015 & -0.0037 & -0.0104 \\ 0 & 0 & 0.1157 & -0.0522 \\ 0 & 0.1098 & -0.0077 & -0.0213 \end{pmatrix}. \tag{5.64}$$

We note that for this uplink scenario, the base station detects the fourth mobile user, which has the spreading sequence corresponds to the fourth column of $\mathbf{S}$, first and the first user last.

If the prescribed SINRs of the four users remain the same in the downlink scenario, the spreading sequences used by the base station are the four columns of the matrix

$$\mathbf{F} = \begin{pmatrix} 17.1936 & -0.2303 & -0.5154 & -1.2796 \\ 0 & 0 & 16.0012 & -6.4449 \\ 0 & 17.0149 & -1.0614 & -2.6352 \end{pmatrix}. \tag{5.65}$$

In this case, the base station applies the dirty paper precoder to the first mobile user first and the last user last. Note that the columns of $\mathbf{F}$ and $\mathbf{W}$ in (5.64) are the same up to a scaling factor. Moreover, $\mathrm{tr}(\mathbf{F}\mathbf{F}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{S}\mathbf{S}^{\mathrm{T}}) = 892.7274$, which means that the power consumed in the base station equals to the overall power used by the four mobile users. At the mobile end, the users use the nulling vectors

$$\mathbf{S} = \begin{pmatrix} 10.0000 & -12.0745 & -6.4974 & -3.0926 \\ 0 & 0 & 7.4138 & -15.5760 \\ 0 & 8.8312 & -13.3801 & -6.3686 \end{pmatrix}, \tag{5.66}$$

which are exactly the spreading sequences used in the uplink scenario. Scaling the output signals may be necessary at the mobile ends for the subsequent dirty paper decoder. But the signal scaling does not influence the output SINR.

If zeros are not allowed in the spreading sequences, we can left multiply $\mathbf{S}$ and $\mathbf{W}$ a 3×3 orthogonal matrix to eliminate the zeros in $\mathbf{S}$.

5.5.5 Further Remarks

The TCD scheme, which was originally motivated by MIMO transceiver designs, turns out to be similar to the scheme of [56] in several aspects. Both schemes are based on the nonlinear decision feedback operations. Hence both are optimal in terms of maximizing the channel throughput and minimizing the overall input power. Both the GTD algorithm, on which the TCD scheme is based, and the construction of the Hermitian matrix with prescribed eigenvalues and Cholesky values as done in [56] rely on the Weyl-Horn theorem. However, our TCD scheme enjoys several remarkable advantages over the scheme of [56]. First, note that if we obtain the GTD $\mathbf{H} = \mathbf{QRP}^*$, where $\mathbf{R}$ has the prescribed diagonal elements, then it follows immediately that $\mathbf{A} \triangleq \mathbf{P}^*\mathbf{H}^*\mathbf{HP} = \mathbf{RR}^*$ is the desired Cholesky decomposition. However, the information associated with $\mathbf{Q}$ is lost in the Cholesky decomposition. Hence the nulling vectors used at the receivers of [56] cannot be calculated explicitly as our TCD does (cf. (5.27)). Furthermore, the correlation matrix $\mathbf{A}$ is only an intermediate result. To get the CDMA sequences, one has to decompose $\mathbf{A} = \mathbf{RR}^*$ explicitly. The TCD scheme, however, can be used to obtain both the precoder (CDMA sequences), which are the columns of $\mathbf{P}$, and the equalizer from $\mathbf{Q}$ simultaneously. Second, our TCD scheme is a solution to the more general MIMO transceiver design problem. The Cholesky decomposition algorithm provided in Appendix C of [56] is only applicable to the scenario where the singular values are only of two values. Hence it is not applicable to the general design of MIMO transceivers. The more general Cholesky factorization algorithm suggested in the proofs is computationally far less efficient. Third, the TCD scheme has two implementation forms, i.e., TCD-VBLAST and TCD-DP, which makes it applicable to both the downlink and uplink scenarios. Finally, the TCD scheme provides insights that identify the CDMA sequence design problem as special cases of the MIMO transceiver design.

5.6 Conclusions

Based on the recently developed GTD matrix decomposition algorithm, we have proposed the TCD scheme utilizing the CSIT and CSIR. TCD can be used to decompose a MIMO channel into multiple subchannels with prescribed capacities. The TCD scheme has two implementation forms. One is the combination of a linear precoder and a minimum mean-squared-error VBLAST (MMSE-VBLAST) detector, which is referred to as TCD-VBLAST, and the other includes a dirty paper (DP) precoder and a linear equalizer followed by a DP decoder, which we refer to as TCD-DP. Both forms of TCD

are computationally very efficient. We have also determined the subchannel capacity region such that a capacity lossless decomposition is possible. The applications of the TCD scheme for MIMO communications with QoS constraints have been investigated. We have also identified the problems of designing precoders for OFDM communications and designing CDMA sequences as special cases in the unifying framework of MIMO transceiver designs. In particular, we have shown that the CDMA sequence design problem in the uplink and downlink scenarios can be solved using TCD-VBLAST and TCD-DP, respectively.

Appendix A
Proof of Theorem 5.4.1

Observe that for $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$, we have

$$\text{tr}\,(\mathbf{F}\mathbf{F}^*) = \sum_{i=1}^{K} \phi_i \quad \text{and} \quad \mathbf{H}\mathbf{F} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Phi}^{1/2}. \tag{5.67}$$

Hence, $\lambda_{HF,i}^2 = \lambda_{H,i}^2 \phi_i$ for $1 \le i \le K$, and

$$\lambda_{G_a,i} = \begin{cases} \sqrt{1 + \lambda_{H,i}^2 \phi_i}, & 1 \le i \le K, \\ 1, & i > K. \end{cases}$$

Since $1+\rho_i \ge 1$, the last $L-K$ inequalities in the multiplicative majorization condition in (5.46) are implied by the single equality constraint in (5.47). Hence, the problem (5.46) reduces to (5.47) where $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$, which gives an upper bound for the minimum in (5.46).

Let $\mathbf{F} = \mathbf{U}_F \mathbf{\Phi}^{1/2} \mathbf{\Omega}^{\mathrm{T}}$ denote the singular value decomposition for any given $\mathbf{F} \in \mathbb{C}^{M_t \times L}$. Once again, tr $(\mathbf{F}\mathbf{F}^*)$ is given by the sum in (5.67). By [60, Theorem 3.3.14], the singular values of the product $\mathbf{H}\mathbf{F}$ of two matrices are multiplicatively majorized by the product of the singular values of $\mathbf{H}$ and $\mathbf{F}$:

$$\prod_{i=1}^{k} \lambda_{H,i}^2 \phi_i \ge \prod_{i=1}^{k} \lambda_{HF,i}^2, \quad 1 \le k \le K. \tag{5.68}$$

Taking log's, we have

$$\sum_{i=1}^{k} \log(\lambda_{H,i}^2 \phi_i) \ge \sum_{i=1}^{k} \log(\lambda_{HF,i}^2), \quad 1 \le k \le K. \tag{5.69}$$

By [60, Lemma 3.3.8]) and (5.69),

$$\sum_{i=1}^{k} f(\log(\lambda_{H,i}^2\phi_i)) \geq \sum_{i=1}^{k} f(\log(\lambda_{HF,i}^2)), \quad 1 \leq k \leq K, \tag{5.70}$$

whenever f is a real-valued, increasing convex function. The function $f(t) = \log(e^t + 1)$ is convex since its second derivative is positive. Making this choice for f in (5.70) and exponentiating both sides, we obtain:

$$\prod_{i=1}^{k}(\lambda_{H,i}^2\phi_i + 1) \geq \prod_{i=1}^{k}(\lambda_{HF,i}^2 + 1), \quad 1 \leq k \leq K. \tag{5.71}$$

Since $\mathbf{F}$ is feasible in (5.46),

$$\begin{aligned} \prod_{i=1}^{k}(\lambda_{HF,i}^2 + 1) &\geq \prod_{i=1}^{k}(\rho_i + 1), \quad 1 \leq k < K, \\ \prod_{i=1}^{K}(\lambda_{HF,i}^2 + 1) &= \prod_{i=1}^{L}(\rho_i + 1). \end{aligned}$$

Combining this with (5.71), we get

$$\begin{aligned} \prod_{i=1}^{k}(\lambda_{H,i}^2\phi_i + 1) &\geq \prod_{i=1}^{k}(\rho_i + 1), \quad 1 \leq k < K, \\ \prod_{i=1}^{K}(\lambda_{H,i}^2\phi_i + 1) &\geq \prod_{i=1}^{L}(\rho_i + 1). \end{aligned} \tag{5.72}$$

Since $\lambda_{H,i}^2\phi_i + 1$ is the square of the i-th singular value of the augmented matrix $\mathbf{G}_a$ corresponding to the choice $\mathbf{V}\mathbf{\Phi}^{1/2}$, we conclude that $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$ satisfies all the inequality constraints in (5.46). If the inequality (5.72) is strict, then ϕ_K should be decreased in order to satisfy the equality constraint in (5.47). Since decreasing ϕ_K only lowers $\mathrm{tr}(\mathbf{F}\mathbf{F}^*)$, we deduce that the minimum in (5.46) is achieved by a matrix of the form $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$. If $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}$ is optimal in (5.46), then so is $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ whenever $\mathbf{\Omega}$ has K orthonormal columns (since the constraints are satisfied and the value of the cost does not change). We now make the choice for $\mathbf{\Omega}$ given in Theorem 5.3.1. That is, if $\mathbf{Q}\mathbf{R}_{G_a}\mathbf{P}^{\mathsf{T}}$ is the GTD of $\mathbf{\Lambda}_{G_a}$ in (5.19) where $\mathbf{\Phi}$ is a solution of (5.47), then $\mathbf{\Omega}$ is the matrix formed by the first K columns of $\mathbf{P}^{\mathsf{T}}$. For this choice of $\mathbf{\Omega}$, the constraints of (5.45) are satisfied. As noted earlier, the minimum in (5.45) can be no smaller than the minimum in (5.46). Since this choice for $\mathbf{F}$ yields the same cost in both (5.45) and (5.46), we conclude that $\mathbf{F} = \mathbf{V}\mathbf{\Phi}^{1/2}\mathbf{\Omega}^{\mathsf{T}}$ is optimal in (5.45).

Appendix B
Proof of Lemma 5.4.3

First suppose that $\gamma_l \geq 1/\lambda_l^2$. By the arithmetic/geometric mean inequality, the problem

$$\min \sum_{i=1}^{l} \psi_i \quad \text{subject to} \quad \prod_{i=1}^{l} \psi_i \geq \prod_{i=1}^{l} \beta_i, \quad \boldsymbol{\psi} \geq \mathbf{0}, \tag{5.73}$$

has the solution $\psi_i = \gamma_l$, $1 \leq i \leq l$. Since $\lambda_{H,i}$ is a decreasing function of i and $\gamma_l \geq 1/\lambda_{H,i}^2$, we conclude that $\psi_i = \gamma_l$ satisfies the constraints $\psi_i \geq 1/\lambda_{H,i}^2$ for $1 \leq i \leq l$. Since l attains the maximum in (5.53),

$$\gamma_l^k \geq \prod_{i=1}^{k} \beta_k$$

for all $k \leq l$. Hence, by taking $\psi_i = \gamma_l$ for $1 \leq i \leq l$, the first l inequalities in (5.49) are satisfied, with equality for $k = l$, and the first l lower bound constraints $\psi_i \geq 1/\lambda_{H,i}^2$ are satisfied.

Let $\boldsymbol{\psi}^*$ denote any optimal solution of (5.49). If

$$\prod_{i=1}^{l} \psi_i^* = \prod_{i=1}^{l} \beta_i, \tag{5.74}$$

then by the unique optimality of $\psi_i = \gamma_l$, $1 \leq i \leq l$, in (5.73), and by the fact that the inequality constraints in (5.49) are satisfied for $k \in [1, l]$, we conclude that $\psi_i^* = \gamma_l$ for all $i \in [1, l]$. On the other hand, suppose that

$$\prod_{i=1}^{l} \psi_i^* > \prod_{i=1}^{l} \beta_l = \gamma_l^l. \tag{5.75}$$

We show below that this leads to a contradiction; consequently, (5.74) holds and $\psi_i^* = \gamma_l$ for $i \in [1, l]$.

Define the quantity

$$\gamma_* = \left(\prod_{i=1}^{l} \psi_i^* \right)^{1/l}.$$

By (5.75) $\gamma_* > \gamma_l$. Again, by the arithmetic/geometric mean inequality, the solution of the problem

$$\min \sum_{i=1}^{l} \psi_i \quad \text{subject to} \quad \prod_{i=1}^{l} \psi_i \geq \gamma_*^l, \quad \boldsymbol{\psi} \geq \mathbf{0}, \tag{5.76}$$

is $\psi_i = \gamma_*$ for $i \in [1, l]$. By (5.75), $\gamma_* > \gamma_l$ and $\boldsymbol{\psi}$ satisfies the inequality constraints in (5.49) for $k \in [1, l]$.

Let M be the first index with the property that

$$\prod_{i=1}^{M} \psi_i^* = \prod_{i=1}^{M} \beta_i. \tag{5.77}$$

Such an index exists since $\boldsymbol{\psi}^*$ is optimal, which implies that

$$\prod_{i=1}^{K} \psi_i^* = \prod_{i=1}^{K} \beta_i.$$

First, suppose that $M \leq j$, where j is the break point given in Lemma 5.4.2. By complementary slackness, $\mu_i = 0$ and $\psi_i^* - \psi_{i+1}^* = \mu_i$ for $1 \leq i < M$. We conclude that $\psi_i = \gamma_*$ for $1 \leq i \leq M$. By (5.77) we have

$$\gamma_*^M = \prod_{i=1}^{M} \beta_i.$$

It follows that

$$\left(\prod_{i=1}^{M} \beta_i \right)^{1/M} = \gamma_* > \gamma_l,$$

which contradicts the fact that l achieves the maximum in (5.53).

In the case $M > j$, we have $\psi_i = \gamma_*$ for $1 \leq i \leq j$. Again, this follows by complementary slackness. However, we need to stop when $i = j$ since the lower bound constraints become active for $i > j$. In Lemma 5.4.2, we show that $\psi_i^* \geq \psi_j^* = \gamma_*$ for $i \geq j$. Consequently, we have

$$\prod_{i=1}^{M} \beta_i = \prod_{i=1}^{M} \psi_i^* \geq \gamma_*^M > \gamma_l^M.$$

Again, this contradicts the fact that l achieves the maximum in (5.53). This completes the analysis of the case where $\gamma_l \geq 1/\lambda_{H,l}^2$.

Now consider the case $\gamma_l < 1/\lambda_{H,l}^2$. By the definition of γ_l, we have

$$\gamma_l \geq \left(\prod_{i=1}^{K} \beta_i \right)^{1/K} \quad \text{or} \quad \gamma_l^K \geq \prod_{i=1}^{K} \beta_i. \tag{5.78}$$

If j is the break point described in Lemma 5.4.2, then $\psi_i^* \geq \psi_j^*$ for all i; it follows that

$$\prod_{i=1}^{K} \psi_i^* \geq \left(\psi_j^* \right)^K. \tag{5.79}$$

Since the product of the components of $\boldsymbol{\psi}^\star$ is equal to the product of the components of $\boldsymbol{\beta}$, from (5.78) and (5.79) we get

$$\gamma_l^K \geq \prod_{i=1}^{K} \beta_i = \prod_{i=1}^{K} \psi_i^\star \geq \left(\psi_j^\star\right)^K .$$

Hence, $\gamma_l \geq \psi_j^\star \geq 1/\lambda_{H,j}^2 \geq 1/\lambda_{H,i}^2$ for all $i \leq j$. In particular, if $l \leq j$, then $\gamma_l \geq 1/\lambda_{H,l}^2$, or, $l > j$ when $\gamma_l < 1/\lambda_{H,l}^2$. As a consequence, $\psi_l^\star = \frac{1}{\lambda_{H,l}^2}$.

CHAPTER 6
NOVEL MATRIX DECOMPOSITIONS

6.1 Introduction

Given a complex matrix $\mathbf{H}$, we consider the decomposition $\mathbf{H} = \mathbf{QRP}^*$, where $\mathbf{R}$ is upper triangular and $\mathbf{Q}$ and $\mathbf{P}$ have orthonormal columns. Special instances of this decomposition are

(a) the singular value decomposition (SVD) [61, 62]

$$\mathbf{H} = \mathbf{V\Sigma W}^*,$$

where $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values on the diagonal,

(b) the Schur decomposition [63]

$$\mathbf{H} = \mathbf{QUQ}^*,$$

where $\mathbf{U}$ is an upper triangular matrix with the eigenvalues of $\mathbf{H}$ on the diagonal,

(c) the QR decomposition where $\mathbf{P} = \mathbf{I}$.

In this chapter, we will introduce two novel matrix decompositions, i.e., the geometric mean decomposition (GMD) and the generalized triangular decomposition (GTD). As we introduced before, the GMD scheme and the UCD scheme are based on the GMD matrix decomposition algorithm, and the TCD is based on the GTD algorithm. The results of this chapter are motivated by the applications of designing MIMO transceiver. Interesting, these results turn out to be also useful to the numerical analysis community.

6.2 Geometric Mean Decomposition

In this section, we present a new unitary decomposition which call the *geometric mean decomposition* or GMD. Given a rank K matrix $\mathbf{H} \in \mathbb{C}^{m \times n}$, it is expressed in the form $\mathbf{QRP}^*$ where $\mathbf{P}$ and $\mathbf{Q}$ have orthonormal columns, and $\mathbf{R} \in \mathbb{R}^{K \times K}$ is a real upper triangular matrix with diagonal elements all equal to the geometric mean of the positive singular values:

$$r_{ii} = \bar{\sigma} = \left(\prod_{\sigma_j > 0} \sigma_j \right)^{1/K}, \quad 1 \leq i \leq K.$$

Here the σ_j are the singular values of $\mathbf{H}$, and $\bar{\sigma}$ is the geometric mean of the positive singular values. Thus $\mathbf{R}$ is upper triangular and the nonzero diagonal elements are the geometric mean of the positive singular values.

We were led to this decomposition when trying to optimize the performance of multiple-input multiple-output (MIMO) systems. However, this decomposition has arisen recently in several other applications. In [64, Prob. 26.3] Higham proposed the following problem:

> Develop an efficient algorithm for computing a unit upper triangular matrix with prescribed singular values σ_i, $1 \leq i \leq K$, where the product of the σ_i is 1.

A solution to this problem could be used to construct test matrices with user specified singular values.

The solution of Kosowski and Smoktunowicz [65] starts with the diagonal matrix Σ, with i-th diagonal element σ_i, and applies a series of 2 by 2 orthogonal transformations to obtain a unit triangular matrix. The complexity of their algorithm is $O(K^2)$. Thus the solution given in [65] amounts to the statement

$$\mathbf{Q}_0^T \Sigma \mathbf{P}_0 = \mathbf{R}, \tag{6.1}$$

where $\mathbf{R}$ is unit upper triangular.

For general $\mathbf{\Sigma}$, where the product of the σ_i is not necessarily 1, one can multiply $\mathbf{\Sigma}$ by the scaling matrix $\bar{\sigma}^{-1}\mathbf{I}$, apply (6.1), then multiply by $\bar{\sigma}$ to obtain the GMD of $\mathbf{\Sigma}$. And for a general matrix $\mathbf{H}$, the singular value decomposition $\mathbf{H} = \mathbf{V}\mathbf{\Sigma}\mathbf{W}^*$ and (6.1) combine to give the $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^*$ where

$$\mathbf{Q} = \mathbf{V}\mathbf{Q}_0 \quad \text{and} \quad \mathbf{P} = \mathbf{W}\mathbf{P}_0.$$

According to (3.11), we consider the problem of choosing $\mathbf{Q}$ and $\mathbf{P}$ to maximize the minimum of the r_{ii}:

$$\begin{aligned} &\max_{\mathbf{Q},\mathbf{P}} \ \min \ \{r_{ii} : 1 \leq i \leq K\} \\ &\text{subject to} \quad \mathbf{Q}\mathbf{R}\mathbf{P}^* = \mathbf{H}, \ \mathbf{Q}^*\mathbf{Q} = \mathbf{I}, \ \mathbf{P}^*\mathbf{P} = \mathbf{I}, \\ &\qquad\qquad r_{ij} = 0 \text{ for } i > j, \ \mathbf{R} \in \mathbb{R}^{K \times K}, \end{aligned} \tag{6.2}$$

where K is the rank of $\mathbf{H}$. Since the GMD of $\mathbf{H}$ is feasible in (6.2), we conclude that the GMD yields the optimal solution to (6.2).

6.2.1 Generalized Maximin Properties

We consider the following problem:

$$\begin{aligned} \max_{\mathbf{F},\mathbf{G}} \ & \min \ \{u_{ii} : 1 \le i \le K\} \\ \text{subject to} \ & \mathbf{GUF}^* = \mathbf{H}, \ u_{ij} = 0 \text{ for } i > j, \ \mathbf{U} \in \mathbb{R}^{K \times K}, \\ & u_{ii} > 0, \ 1 \le i \le K, \\ & \text{tr}\,((\mathbf{G}^*\mathbf{G})^{-1}) \le p_1, \ \text{tr}\,((\mathbf{F}^*\mathbf{F})^{-1}) \le p_2. \end{aligned} \tag{6.3}$$

If $p_1 = p_2 = K$, then any $\mathbf{Q}$ and $\mathbf{P}$ feasible in (6.2) are feasible in (6.3). Hence, the problem (6.3) is less constrained than the problem (6.2) since the set of feasible matrices has been enlarged. Nonetheless, we now show that the solution to this relaxed problem is the same as the solution of the more constrained problem (6.2).

Theorem 6.2.1 *If $\mathbf{H} \in \mathbb{C}^{m \times n}$ has rank K, then a solution of* (6.3) *is given by*

$$\mathbf{G} = \mathbf{Q}\sqrt{\frac{K}{p_1}}, \quad \mathbf{U} = \left(\frac{\sqrt{p_1 p_2}}{K}\right)\mathbf{R}, \quad \textit{and} \quad \mathbf{F} = \mathbf{P}\sqrt{\frac{K}{p_2}},$$

where $\mathbf{QRP}^*$ is the GMD of $\mathbf{H}$.

Proof: Let $\mathbf{V\Sigma W}^*$ be the singular value decomposition of $\mathbf{H}$, where $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ contains the K positive singular values of $\mathbf{H}$ on the diagonal. If $\mathbf{F}$ and $\mathbf{G}$ satisfy the constraints of (6.3), then we have

$$\mathbf{H} = \mathbf{V\Sigma W}^* = \mathbf{GUF}^*.$$

The column space of $\mathbf{GUF}^*$ is contained in the column space of $\mathbf{G}$. Since $\mathbf{G}$ has K columns, the dimension of the column space is at most K. Since $\mathbf{GUF}^* = \mathbf{H}$ has rank K, the column space of $\mathbf{G}$ must coincide with the column space of $\mathbf{H}$, which is equal to the column space of $\mathbf{V}$. Hence, there exists a K by K invertible matrix $\mathbf{A}$ such that

$$\mathbf{G} = \mathbf{VA}. \tag{6.4}$$

In the same fashion, the column space of $\mathbf{F}$ must coincide with the column space of $\mathbf{H}^*$, which is equal to the column space of $\mathbf{W}$. And there exists a K by K invertible matrix $\mathbf{B}$ such that

$$\mathbf{F} = \mathbf{WB}. \tag{6.5}$$

Combining (6.4) and (6.5) with the identity $\mathbf{GUF}^* = \mathbf{H} = \mathbf{V\Sigma W}^*$ gives

$$\mathbf{AUB}^* = \mathbf{\Sigma}.$$

It follows that

$$\det(\mathbf{\Sigma}^*\mathbf{\Sigma}) = \det(\mathbf{BU}^*\mathbf{A}^*\mathbf{AUB}^*) = \det(\mathbf{A}^*\mathbf{A})\det(\mathbf{B}^*\mathbf{B})\prod_{i=1}^{K}|u_{ii}|^2,$$

which gives

$$\min_{1\le i\le K}|u_{ii}|^{2K} \le \prod_{i=1}^{K}|u_{ii}|^2 = \det(\mathbf{\Sigma}^*\mathbf{\Sigma})\det(\mathbf{A}^*\mathbf{A})^{-1}\det(\mathbf{B}^*\mathbf{B})^{-1}. \tag{6.6}$$

By the constraints of (6.3), we have

$$\begin{aligned}\operatorname{tr}((\mathbf{G}^*\mathbf{G})^{-1}) &= \operatorname{tr}((\mathbf{A}^*\mathbf{A})^{-1}) \le p_1,\\ \operatorname{tr}((\mathbf{F}^*\mathbf{F})^{-1}) &= \operatorname{tr}((\mathbf{B}^*\mathbf{B})^{-1}) \le p_2.\end{aligned}$$

By the geometric mean inequality and the fact that the determinant (trace) of a matrix is the product (sum) of the eigenvalues, a K by K Hermitian positive semidefinite matrix $\mathbf{S}$ satisfies

$$\det(\mathbf{S}) \le \left(\frac{\operatorname{tr}(\mathbf{S})}{K}\right)^K.$$

Using these bounds for the determinant and the trace in (6.6), we have

$$\min_{1\le i\le K}|u_{ii}| \le \frac{\bar{\sigma}\sqrt{p_1 p_2}}{K}. \tag{6.7}$$

Finally, it can be verified that for the choices of $\mathbf{G}$, $\mathbf{U}$, and $\mathbf{F}$ given in the statement of the theorem, the inequality (6.7) is an equality. □ ■

6.2.2 Implementation Based on Initial SVD

We now give an algorithm for evaluating the GMD that starts with the singular value decomposition $\mathbf{H} = \mathbf{V\Sigma W}^*$. The algorithm generates a sequence of upper triangular matrices $\mathbf{R}^{(L)}$, $1 \le L < K$, with $\mathbf{R}^{(1)} = \mathbf{\Sigma}$. Each matrix $\mathbf{R}^{(L)}$ has the following properties:

(a) $r_{ij}^{(L)} = 0$ when $i > j$ or $j > \max\{L, i\}$.

(b) $r_{ii}^{(L)} = \bar{\sigma}$ for all $i < L$, and the geometric mean of $r_{ii}^{(L)}$, $L \le i \le K$, is $\bar{\sigma}$.

We express $\mathbf{R}^{(k+1)} = \mathbf{Q}_k\mathbf{R}^{(k)}\mathbf{P}_k$ where $\mathbf{Q}_k$ and $\mathbf{P}_k$ are orthogonal for each k.

These orthogonal matrices are constructed using a symmetric permutation and a pair of Givens rotations. Suppose that $\mathbf{R}^{(k)}$ satisfies (a) and (b). If $r_{kk}^{(k)} \ge \bar{\sigma}$, then let

$\mathbf{\Pi}$ be a permutation matrix with the property that $\mathbf{\Pi R}^{(k)}\mathbf{\Pi}$ exchanges the $(k+1)$-st diagonal element of $\mathbf{R}^{(k)}$ with any element r_{pp}, $p > k$, for which $r_{pp} \leq \bar{\sigma}$. If $r_{kk}^{(k)} < \bar{\sigma}$, then let $\mathbf{\Pi}$ be chosen to exchange the $(k+1)$-st diagonal element with any element r_{pp}, $p > k$, for which $r_{pp} \geq \bar{\sigma}$. Let $\delta_1 = r_{kk}^{(k)}$ and $\delta_2 = r_{pp}^{(k)}$ denote the new diagonal elements at locations k and $k+1$ associated with the permuted matrix $\mathbf{\Pi R}^{(k)}\mathbf{\Pi}$.

Next, we construct orthogonal matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ by modifying the elements in the identity matrix that lie at the intersection of rows k and $k+1$ and columns k and $k+1$. We multiply the permuted matrix $\mathbf{\Pi R}^{(k)}\mathbf{\Pi}$ on the left by $\mathbf{G}_2^{\mathrm{T}}$ and on the right by $\mathbf{G}_1$. These multiplications will change the elements in the 2 by 2 submatrix at the intersection of rows k and $k+1$ with columns k and $k+1$. Our choice for the elements of $\mathbf{G}_1$ and $\mathbf{G}_2$ is shown below, where we focus on the relevant 2 by 2 submatrices of $\mathbf{G}_2^{\mathrm{T}}$, $\mathbf{\Pi R}^{(k)}\mathbf{\Pi}$, and $\mathbf{G}_1$:

$$\underset{(\mathbf{G}_2^{\mathrm{T}})}{\bar{\sigma}^{-1}\begin{bmatrix} c\delta_1 & s\delta_2 \\ -s\delta_2 & c\delta_1 \end{bmatrix}} \quad \underset{(\mathbf{\Pi R}^{(k)}\mathbf{\Pi})}{\begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}} \quad \underset{(\mathbf{G}_1)}{\begin{bmatrix} c & -s \\ s & c \end{bmatrix}} = \underset{(\mathbf{R}^{(k+1)})}{\begin{bmatrix} \bar{\sigma} & x \\ 0 & y \end{bmatrix}} \tag{6.8}$$

If $\delta_1 = \delta_2 = \bar{\sigma}$, we take $c = 1$ and $s = 0$; if $\delta_1 \neq \delta_2$, we take

$$c = \sqrt{\frac{\bar{\sigma}^2 - \delta_2^2}{\delta_1^2 - \delta_2^2}} \quad \text{and} \quad s = \sqrt{1 - c^2}. \tag{6.9}$$

In either case,

$$x = \frac{sc(\delta_2^2 - \delta_1^2)r_k}{\bar{\sigma}} \quad \text{and} \quad y = \frac{\delta_1\delta_2}{\bar{\sigma}}. \tag{6.10}$$

Since $\bar{\sigma}$ lies between δ_1 and δ_2, s and c are nonnegative real scalars.

Figure 6–1 depicts the transformation from $\mathbf{R}^{(k)}$ to $\mathbf{G}_2^{\mathrm{T}}\mathbf{\Pi R}^{(k)}\mathbf{\Pi G}_1$. The dashed box is the 2 by 2 submatrix displayed in (6.8). Notice that c and s, defined in (6.9), are real scalars chosen so that

$$c^2 + s^2 = 1 \quad \text{and} \quad (c\delta_1)^2 + (s\delta_2)^2 = \bar{\sigma}^2.$$

With these identities, the validity of (6.8) follows by direct computation. Defining $\mathbf{Q}_k = \mathbf{\Pi G}_2$ and $\mathbf{P}_k = \mathbf{\Pi G}_1$, we set

$$\mathbf{R}^{(k+1)} = \mathbf{Q}_k^{\mathrm{T}}\mathbf{R}^{(k)}\mathbf{P}_k. \tag{6.11}$$

It follows from Figure 6–1, (6.8), and the identity $\det(\mathbf{R}^{(k+1)}) = \det(\mathbf{R}^{(k)})$, that (a) and (b) hold for $L = k+1$. Thus there exists a real upper triangular matrix $\mathbf{R}^{(K)}$, with

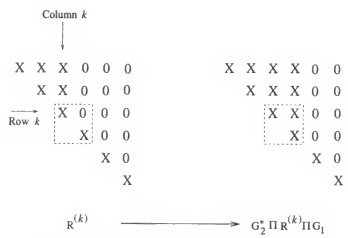

Figure 6–1: The operation displayed in (6.8)

$\bar{\sigma}$ on the diagonal, and unitary matrices $\mathbf{Q}_i$ and $\mathbf{P}_i$, $i = 1, 2, \ldots, K-1$, such that

$$\mathbf{R}^{(K)} = (\mathbf{Q}_{k-1}^T \ldots \mathbf{Q}_2^T \mathbf{Q}_1^T) \Sigma (\mathbf{P}_1 \mathbf{P}_2 \ldots \mathbf{P}_{k-1}).$$

Combining this identity with the singular value decomposition, we obtain $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^*$ where

$$\mathbf{Q} = \mathbf{V} \left(\prod_{i=1}^{K-1} \mathbf{Q}_i \right), \quad \mathbf{R} = \mathbf{R}^{(K)}, \quad \text{and} \quad \mathbf{P} = \mathbf{W} \left(\prod_{i=1}^{K-1} \mathbf{P}_i \right).$$

In summary, our algorithm for computing the GMD, based on an initial SVD, is the following:

1. Let $\mathbf{H} = \mathbf{V}\Sigma\mathbf{W}^*$ be the singular value decomposition of $\mathbf{H}$, and initialize $\mathbf{Q} = \mathbf{V}$, $\mathbf{P} = \mathbf{W}$, $\mathbf{R} = \Sigma$, and $k = 1$.
2. If $r_{kk} \geq \bar{\sigma}$. choose $p > k$ such that $r_{pp} \leq \bar{\sigma}$. If $r_{kk} < \bar{\sigma}$, choose $p > k$ such that $r_{pp} \geq \bar{\sigma}$. In $\mathbf{R}$, $\mathbf{P}$, and $\mathbf{Q}$, perform the following exchanges:

$$\begin{aligned} r_{k+1,k+1} &\leftrightarrow r_{pp} \\ \mathbf{P}_{:,k+1} &\leftrightarrow \mathbf{P}_{:,p} \\ \mathbf{Q}_{:,k} &\leftrightarrow \mathbf{Q}_{:,p} \end{aligned}$$

3. Construct the matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ shown in (6.8). Replace $\mathbf{R}$ by $\mathbf{G}_2^T \mathbf{R} \mathbf{G}_1$, replace $\mathbf{Q}$ by $\mathbf{Q}\mathbf{G}_2$, and replace $\mathbf{P}$ by $\mathbf{P}\mathbf{G}_1$.
4. If $k = K - 1$, then stop, $\mathbf{Q}\mathbf{R}\mathbf{P}^*$ is the GMD of $\mathbf{H}$. Otherwise, replace k by $k + 1$ and go to step 2.

Given the SVD, this algorithm for the GMD requires $O((m + n)K)$ flops. For comparison, reduction of $\mathbf{H}$ to bidiagonal form by the Golub-Reinsch bidiagonalization

scheme [66, 67, 68], often the first step in the computation of the SVD, requires $O(mnK)$ flops.

6.3 Generalized Triangular Decomposition

In this section, we attempt to generalized decomposition of the form

$$\mathbf{H} = \mathbf{QRP}^*, \tag{6.12}$$

where $\mathbf{R}$ is upper triangular and $\mathbf{Q}$ and $\mathbf{P}$ have orthonormal columns. We will answer the following two questions. First, what is the necessary and sufficient condition that the decomposition of (6.12) exists. Second, how to calculate such a decomposition. Section 6.3.1 and 6.3.2 focus on answering the two questions.

6.3.1 Existence of GTD

The following result is due to Weyl [69] (also see [60, p. 171]):

Theorem 6.3.1 *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ with eigenvalues $\boldsymbol{\lambda}$ and singular values $\boldsymbol{\sigma}$, then $\boldsymbol{\lambda} \preceq \boldsymbol{\sigma}$.*

The following result is due to Horn [70] (also see [60, p. 220]):

Theorem 6.3.2 *If $\mathbf{r} \in \mathbb{C}^n$ and $\boldsymbol{\sigma} \in \mathbb{R}^n$ with $\mathbf{r} \preceq \boldsymbol{\sigma}$, then there exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ with singular values σ_i, $1 \le i \le n$, and with $\mathbf{r}$ on the diagonal of $\mathbf{R}$.*

We now combine Theorems 6.3.1 and 6.3.2 to obtain:

Theorem 6.3.1 *Let $\mathbf{H} \in \mathbb{C}^{m\times n}$ have rank K with singular values $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_K > 0$. There exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K\times K}$ and matrices $\mathbf{Q}$ and $\mathbf{P}$ with orthonormal columns such that $\mathbf{H} = \mathbf{QRP}^*$ if and only if $\mathbf{r} \preceq \boldsymbol{\sigma}$.*

Proof: If $\mathbf{H} = \mathbf{QRP}^*$, then the eigenvalues of $\mathbf{R}$ are its diagonal elements and the singular values of $\mathbf{R}$ coincide with those of $\mathbf{H}$. By Theorem 6.3.1, $\mathbf{r} \preceq \boldsymbol{\sigma}$. Conversely, suppose that $\mathbf{r} \preceq \boldsymbol{\sigma}$. Let $\mathbf{H} = \mathbf{V\Sigma W}^*$ be the singular value decomposition, where $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$. By Theorem 6.3.2, there exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K\times K}$ with the r_i on the diagonal and with singular values σ_i, $1 \le i \le K$. Let $\mathbf{R} = \mathbf{V}_0\boldsymbol{\Sigma}\mathbf{W}_0^*$ be the singular value decomposition of $\mathbf{R}$. Substituting $\boldsymbol{\Sigma} = \mathbf{V}_0^*\mathbf{RW}_0$ in the singular value decomposition for $\mathbf{H}$, we have

$$\mathbf{H} = (\mathbf{VV}_0^*)\mathbf{R}(\mathbf{WW}_0^*)^*.$$

In other words, $\mathbf{H} = \mathbf{QRP}^*$ where $\mathbf{Q} = \mathbf{VV}_0^*$ and $\mathbf{P} = \mathbf{WW}_0^*$. ∎

6.3.2 The GTD Algorithm

Given a matrix $\mathbf{H} \in \mathbb{C}^{m\times n}$ with rank K and with singular values $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_K > 0$, and given a vector $\mathbf{r} \in \mathbb{C}^K$ such that $\mathbf{r} \preceq \boldsymbol{\sigma}$, we now give an algorithm for

computing the decomposition $\mathbf{H} = \mathbf{QRP}^*$. This algorithm for the GTD essentially yields a constructive proof of Theorem 6.3.2.

Let $\mathbf{V\Sigma W}^*$ be the singular value decomposition of $\mathbf{H}$, where $\mathbf{\Sigma}$ is a K by K diagonal matrix with the diagonal containing the positive singular values. We let $\mathbf{R}^{(L)} \in \mathbb{C}^{K\times K}$ denote an upper triangular matrix with the following properties:

(a) $r_{ij}^{(L)} = 0$ when $i > j$ or $j > i \geq L$. In other words, the trailing principal submatrix of $\mathbf{R}^{(L)}$, starting at row L and column L, is diagonal.

(b) If $\mathbf{r}^{(L)}$ denotes the diagonal of $\mathbf{R}^{(L)}$, then the first $L-1$ elements of $\mathbf{r}$ and $\mathbf{r}^{(L)}$ are equal. In other words, the leading diagonal elements of $\mathbf{R}^{(L)}$ match the prescribed leading elements of the vector $\mathbf{r}$.

(c) $\mathbf{r}_{L:K} \preceq \mathbf{r}_{L:K}^{(L)}$, where $\mathbf{r}_{L:K}$ denotes the subvector of $\mathbf{r}$ consisting of components L through K. In other words, the trailing diagonal elements of $\mathbf{R}^{(L)}$ multiplicatively majorize the trailing elements of the prescribed vector $\mathbf{r}$.

Initially, we set $\mathbf{R}^{(1)} = \mathbf{\Sigma}$. Clearly, (a)–(c) hold for $L = 1$. Proceeding by induction, suppose we have generated upper triangular matrices $\mathbf{R}^{(L)}$, $L = 1, 2, \ldots, k$, satisfying (a)–(c), and unitary matrices $\mathbf{Q}_L$ and $\mathbf{P}_L$, such that $\mathbf{R}^{(L+1)} = \mathbf{Q}_L^* \mathbf{R}^{(L)} \mathbf{P}_L$ for $1 \leq L < k$. We now show how to construct unitary matrices $\mathbf{Q}_k$ and $\mathbf{P}_k$ such that $\mathbf{R}^{(k+1)} = \mathbf{Q}_k^* \mathbf{R}^{(k)} \mathbf{P}_k$, where $\mathbf{R}^{(k+1)}$ satisfies (a)–(c) for $L = k+1$.

Let p and q be defined as follows:

$$p = \arg \min_i \{|r_i^{(k)}| : k \leq i \leq K,\ |r_i^{(k)}| \geq |r_k|\}, \tag{6.13}$$

$$q = \arg \max_i \{|r_i^{(k)}| : k \leq i \leq K,\ |r_i^{(k)}| \leq |r_k|,\ i \neq p\}, \tag{6.14}$$

where $r_i^{(k)}$ is the i-th element of $\mathbf{r}^{(k)}$. Since $\mathbf{r}_{k:K} \preceq \mathbf{r}_{k:K}^{(k)}$, there exists p and q satisfying (6.13) and (6.14). Let $\mathbf{\Pi}$ be the matrix corresponding to the symmetric permutation $\mathbf{\Pi}^* \mathbf{R}^{(k)} \mathbf{\Pi}$ which moves the diagonal elements $r_{pp}^{(k)}$ and $r_{qq}^{(k)}$ to the k-th and $(k+1)$-st diagonal positions respectively. Let $\delta_1 = r_{pp}^{(k)}$ and $\delta_2 = r_{qq}^{(k)}$ denote the new diagonal elements at locations k and $k+1$ associated with the permuted matrix $\mathbf{\Pi}^* \mathbf{R}^{(k)} \mathbf{\Pi}$.

Next, we construct unitary matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ by modifying the elements in the identity matrix that lie at the intersection of rows k and $k+1$ and columns k and $k+1$. We multiply the permuted matrix $\mathbf{\Pi}^* \mathbf{R}^{(k)} \mathbf{\Pi}$ on the left by $\mathbf{G}_2^*$ and on the right by $\mathbf{G}_1$. These multiplications will change the elements in the 2 by 2 submatrix at the intersection of rows k and $k+1$ with columns k and $k+1$. Our choice for the elements of $\mathbf{G}_1$ and $\mathbf{G}_2$ is shown below, where we focus on the relevant 2 by 2 submatrices of $\mathbf{G}_2^*$,

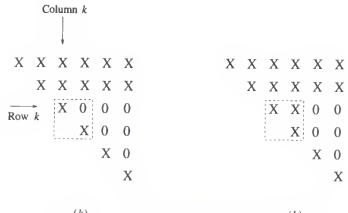

Figure 6–2: The operation displayed in (6.15)

$\Pi^* \mathbf{R}^{(k)} \Pi$, and $\mathbf{G}_1$:

$$\underbrace{\frac{r_k}{|r_k|^2}\begin{bmatrix} c\delta_1^* & s\delta_2^* \\ -s\delta_2 & c\delta_1 \end{bmatrix}}_{(\mathbf{G}_2^*)} \underbrace{\begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}}_{(\Pi^* \mathbf{R}^{(k)} \Pi)} \underbrace{\begin{bmatrix} c & -s \\ s & c \end{bmatrix}}_{(\mathbf{G}_1)} = \underbrace{\begin{bmatrix} r_k & x \\ 0 & y \end{bmatrix}}_{(\mathbf{R}^{(k+1)})} \tag{6.15}$$

If $|\delta_1| = |\delta_2| = |r_k|$, we take $c = 1$ and $s = 0$; if $|\delta_1| \neq |\delta_2|$, we take

$$c = \sqrt{\frac{|r_k|^2 - |\delta_2|^2}{|\delta_1|^2 - |\delta_2|^2}} \quad \text{and} \quad s = \sqrt{1 - c^2}. \tag{6.16}$$

In either case,

$$x = \frac{sc(|\delta_2|^2 - |\delta_1|^2) r_k}{|r_k|^2} \quad \text{and} \quad y = \frac{\delta_1 \delta_2 r_k}{|r_k|^2}. \tag{6.17}$$

$\Pi^* \mathbf{R}^{(k)} \Pi$ to $\mathbf{G}_2^* \Pi^* \mathbf{R}^{(k)} \Pi \mathbf{G}_1$. The dashed box is the 2 by 2 submatrix displayed in (6.15). Notice that c and s, defined in (6.16), are real scalars chosen so that

$$c^2 + s^2 = 1 \quad \text{and} \quad c^2 |\delta_1|^2 + s^2 |\delta_2|^2 = |r_k|^2. \tag{6.18}$$

With these identities, the validity of (6.15) follows by direct computation. By the choice of p and q, we have

$$|\delta_2| \leq |r_k| \leq |\delta_1|. \tag{6.19}$$

If $|\delta_1| \neq |\delta_2|$, it follows from (6.19) that c and s are real nonnegative scalars. It can be checked that the 2 by 2 matrices in (6.15) associated with $\mathbf{G}_1$ and $\mathbf{G}_2^*$ are both unitary. Consequently, both $\mathbf{G}_1$ and $\mathbf{G}_2$ are unitary. We define

$$\mathbf{R}^{(k+1)} = (\Pi \mathbf{G}_2)^* \mathbf{R}^{(k)} (\Pi \mathbf{G}_1) = \mathbf{Q}_k^* \mathbf{R}^{(k)} \mathbf{P}_k,$$

where $\mathbf{Q}_k = \mathbf{\Pi G}_2$ and $\mathbf{P}_k = \mathbf{\Pi G}_1$. By (6.15) and Figure 6–2, $\mathbf{R}^{(k+1)}$ has properties (a) and (b) for $L = k + 1$. Now consider property (c).

We write $\mathbf{a} \sim \mathbf{b}$ if $\mathbf{a}$ and $\mathbf{b}$ are equal after a suitable reordering of the components. Let $\mathbf{a}$, $\mathbf{b}$, $\mathbf{a}^+$, and $\mathbf{b}^+$ be vectors whose components are ordered in decreasing magnitude, and which satisfy

$$\mathbf{a} \sim \mathbf{r}_{k:K}, \quad \mathbf{b} \sim \mathbf{r}^{(k)}_{k:K}, \quad \mathbf{a}^+ \sim \mathbf{r}_{k+1:K}, \quad \text{and} \quad \mathbf{b}^+ \sim \mathbf{r}^{(k+1)}_{k+1:K}. \tag{6.20}$$

Thus a_i is the i-th largest (in magnitude) component of $\mathbf{r}_{k:K}$. By the induction hypothesis, we have $\mathbf{a} \preceq \mathbf{b}$. To establish (c), we need to show that $\mathbf{a}^+ \preceq \mathbf{b}^+$. Let the index s be chosen so that $a_s = r_k$, and let the index t be chosen so that

$$|b_t| \geq |r_k| \geq |b_{t+1}|. \tag{6.21}$$

By the definition of p and q, $r^{(k)}_{pp} = b_t$ and $r^{(k)}_{qq} = b_{t+1}$. As seen in (6.20), $\mathbf{a}^+$ is obtained from $\mathbf{a}$ by deleting $a_s = r_k$. The vector $\mathbf{r}^{(k+1)}$ is obtained from $\mathbf{r}^{(k)}$ by a unitary transformation that changes the value of two elements. In particular, $\mathbf{b}^+$ is obtained from $\mathbf{b}$ by replacing the adjacent pair b_t and b_{t+1} by

$$y = \frac{b_t b_{t+1} r_k}{|r_k|^2}.$$

By (6.21) $|b_t| \geq |y| \geq |b_{t+1}|$. Consequently,

$$b^+_t = y. \tag{6.22}$$

We partition the proof of (c) into 2 cases.

Case 1: $s \leq t$. Since $a^+_i \leq a_i$ for all i, $\mathbf{a} \preceq \mathbf{b}$, and $b_i = b^+_i$ for $1 \leq i < t$, we have

$$\mathbf{a}^+_{1:t-1} \prec \mathbf{a}_{1:t-1} \prec \mathbf{b}_{1:t-1} = \mathbf{b}^+_{1:t-1}. \tag{6.23}$$

For $j > t \geq s$, it follows from the induction hypothesis and the connection between $\mathbf{a}$ and $\mathbf{a}^+$ that

$$|r_k| \prod_{i=1}^{j-1} |a^+_i| = |a_s| \prod_{i=1}^{j-1} |a^+_i| = \prod_{i=1}^{j} |a_i| \leq \prod_{i=1}^{j} |b_i|. \tag{6.24}$$

Since $\mathbf{G}_1$ and $\mathbf{G}_2$ are unitary, the determinant of (6.15) gives

$$|\delta_1 \delta_2| = |r^{(k)}_p r^{(k)}_q| = |b_t b_{t+1}| = |r_k y| = |r_k b^+_t|, \tag{6.25}$$

where the last equality in (6.25) comes from (6.22). Hence, for $j > t$, it follows that

$$\begin{aligned}\prod_{i=1}^{j}|b_i| &= |b_t||b_{t+1}|\left(\prod_{i=1}^{t-1}|b_i|\right)\left(\prod_{i=t+2}^{j}|b_i|\right)\\ &= |r_k||b_t^+|\left(\prod_{i=1}^{t-1}|b_i^+|\right)\left(\prod_{i=t+1}^{j-1}|b_i^+|\right) = |r_k|\prod_{i=1}^{j-1}|b_i^+|. \end{aligned} \tag{6.26}$$

Combining (6.23), (6.24), and (6.26), we have $\mathbf{a}^+ \preceq \mathbf{b}^+$.

Case 2: $s > t$. As before, (6.23) holds. For $t < j < s$, we have

$$\prod_{i=1}^{j}|a_i^+| = \prod_{i=1}^{j}|a_i| \le \prod_{i=1}^{j}|b_i| = |r_k|\prod_{i=1}^{j-1}|b_i^+|,$$

where the first equality comes from the relation $j < s$, the middle inequality is the induction hypothesis, and the last equality is (6.26). Rearranging this gives

$$\left(\frac{|a_j|}{|r_k|}\right)\prod_{i=1}^{j-1}|a_i^+| \le \prod_{i=1}^{j-1}|b_i^+|. \tag{6.27}$$

Since $|a_j|/|r_k| = |a_j|/|a_s| \ge 1$ when $j < s$, we deduce from (6.27) that

$$\mathbf{a}_{1:j-1}^+ \prec \mathbf{b}_{1:j-1}^+$$

when $j < s$. This also holds for $j \ge s$ due to (6.24) and (6.26). This completes the proof of (c).

Hence, there exists an upper triangular matrix $\mathbf{R}^{(K)}$, with $\mathbf{r}_{1:K-1}$ occupying the first $K-1$ diagonal elements, and unitary matrices $\mathbf{Q}_i$ and $\mathbf{P}_i$, $i = 1, 2, \ldots, K-1$, such that

$$\mathbf{R}^{(K)} = (\mathbf{Q}_{k-1}^* \ldots \mathbf{Q}_2^*\mathbf{Q}_1^*)\boldsymbol{\Sigma}(\mathbf{P}_1\mathbf{P}_2 \ldots \mathbf{P}_{k-1}). \tag{6.28}$$

Equating determinants in (6.28) and utilizing the identity $r_i^{(k)} = r_i$ for $1 \le i \le K-1$, we have

$$\prod_{i=1}^{K}|r_i^{(K)}| = \frac{|r_K^{(K)}|}{|r_K|}\left(\prod_{i=1}^{K}|r_i|\right) = \prod_{i=1}^{K}\sigma_i = \prod_{i=1}^{K}|r_i|,$$

where the last equality is due to the assumption $\mathbf{r} \preceq \boldsymbol{\sigma}$. It follows that $|r_K^{(K)}| = |r_K|$. Let $\mathbf{C}$ be the diagonal matrix obtained by replacing the (K,K) element of the identity matrix by $r_K^{(K)}/r_K$. The matrix $\mathbf{C}$ is unitary since $|r_k|/|r_K^{(K)}| = 1$. The matrix

$$\mathbf{R} = \mathbf{C}^*\mathbf{R}^{(K)} \tag{6.29}$$

has diagonal equal to $\mathbf{r}$ due to the choice of $\mathbf{C}$.

Combining (6.28) and (6.29) with the singular value decomposition $\mathbf{H} = \mathbf{V}\mathbf{\Sigma}\mathbf{W}^*$ gives

$$\mathbf{H} = \mathbf{V}\mathbf{Q}_1\mathbf{Q}_2 \ldots \mathbf{Q}_{k-1}\mathbf{C}\mathbf{R}\mathbf{P}_{k-1}^* \ldots \mathbf{P}_2^*\mathbf{P}_1^*\mathbf{W}^*.$$

Hence, we have obtained the GTD with

$$\mathbf{Q} = \mathbf{V}\left(\prod_{i=1}^{K-1}\mathbf{Q}_i\right)\mathbf{C} \quad \text{and} \quad \mathbf{P} = \mathbf{W}\left(\prod_{i=1}^{K-1}\mathbf{P}_i\right).$$

Finally, note that if $\mathbf{r}$ is real, then $\mathbf{G}_1$ and $\mathbf{G}_2$ are real, which implies $\mathbf{R}$ is real.

We summarize the steps of the GTD algorithm as follows. To make it easier to distinguish between the elements of the matrix $\mathbf{R}$ and the elements of the given diagonal vector $\mathbf{r}$, we use R_{ij} to denote the (i,j) element of $\mathbf{R}$ and r_i to denote the i-th element of $\mathbf{r}$.

1. Let $\mathbf{H} = \mathbf{V}\mathbf{\Sigma}\mathbf{W}^*$ be the singular value decomposition of $\mathbf{H}$, and suppose we are given $\mathbf{r} \in \mathbb{C}^K$ with $\mathbf{r} \preceq \boldsymbol{\sigma}$. Initialize $\mathbf{Q} = \mathbf{V}$, $\mathbf{P} = \mathbf{W}$, $\mathbf{R} = \mathbf{\Sigma}$, and $k = 1$.
2. Let p and q be defined as follows:

$$\begin{aligned} p &= \arg\ \min_i\{|R_{ii}| : k \le i \le K,\ |R_{ii}| \ge |r_k|\}, \\ q &= \arg\ \max_i\{|R_{ii}| : k \le i \le K,\ |R_{ii}| \le |r_k|,\ i \ne p\}. \end{aligned}$$

   In $\mathbf{R}$, $\mathbf{P}$, and $\mathbf{Q}$, perform the following exchanges:

$$\begin{aligned} (R_{kk}, R_{k+1,k+1}) &\leftrightarrow (R_{pp}, R_{qq}) \\ (\mathbf{R}_{1:k-1,k}, \mathbf{R}_{1:k-1,k+1}) &\leftrightarrow (\mathbf{R}_{1:k-1,p}, \mathbf{R}_{1:k-1,q}) \\ (\mathbf{P}_{:,k}, \mathbf{P}_{:,k+1}) &\leftrightarrow (\mathbf{P}_{:,p}, \mathbf{P}_{:,q}) \\ (\mathbf{Q}_{:,k}, \mathbf{Q}_{:,k+1}) &\leftrightarrow (\mathbf{Q}_{:,p}, \mathbf{Q}_{:,q}) \end{aligned}$$

3. Construct the matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ shown in (6.15). Replace $\mathbf{R}$ by $\mathbf{G}_2^*\mathbf{R}\mathbf{G}_1$, replace $\mathbf{Q}$ by $\mathbf{Q}\mathbf{G}_2$, and replace $\mathbf{P}$ by $\mathbf{P}\mathbf{G}_1$.
4. If $k = K - 1$, then go to step 5. Otherwise, replace k by $k + 1$ and go to step 2.
5. Multiply column K of $\mathbf{Q}$ by R_{KK}/r_K; replace R_{KK} by r_K. The product $\mathbf{Q}\mathbf{R}\mathbf{P}^*$ is the GTD of $\mathbf{H}$ based on $\mathbf{r}$.

A Matlab implementation of our GTD algorithm appears in the Appendix. Given the SVD, this algorithm for the GTD requires $O((m+n)K)$ flops. For comparison, reduction of $\mathbf{H}$ to bidiagonal form by the Golub-Reinsch bidiagonalization scheme [66, 67, 68], often the first step in the computation of the SVD, requires $O(mnK)$ flops.

	Time		σ error		λ error	
Dimension	SVD_EIG	GTD	SVD_EIG	GTD	SVD_EIG	GTD
100	0.61	0.20	9.8e−15	1.0e−14	3.3e−14	0
200	2.24	0.38	2.0e−14	1.7e−14	5.9e−13	0
400	13.84	0.86	6.8e−14	3.7e−14	3.3e−13	0
800	97.50	2.30	9.8e−14	7.0e−14	1.5e−10	0
1200	317.83	5.67	1.1e−13	1.3e−13	1.5e−9	0
1600	746.77	10.77	3.2e−13	1.5e−13	7.7e−4	0

Table 6–1: Comparison of SVD_EIG and GTD for inverse eigenvalue problems (CPU time in seconds, singular value and eigenvalue errors in sup-norm)

6.3.3 Inverse Eigenvalue Problem

In [71] Chu presents a recursive procedure for constructing matrices with prescribed eigenvalues and singular values. His algorithm, which he calls SVD_EIG, is based on Horn's divide and conquer proof of the sufficiency of Weyl's product inequalities. In general, the output of SVD_EIG is not upper triangular. Consequently, this routine could not be used to generate the GTD. Chu notes that to achieve an upper triangular matrix would require an algorithm "one order more expensive than the divide-and-conquer algorithm".

Given a vector of singular values $\boldsymbol{\sigma} \in \mathbb{R}^n$ and a vector of eigenvalues $\boldsymbol{\lambda} \in \mathbb{C}^n$, with $\boldsymbol{\lambda} \preceq \boldsymbol{\sigma}$, we can use the GTD to generate a matrix $\mathbf{R}$ with $\boldsymbol{\lambda}$ on the diagonal and with singular values $\boldsymbol{\sigma}$. In this section, we compare the solution to the inverse eigenvalue problem provided by the GTD to Chu's algorithm. In our initial experimentation, we discovered that the algorithm of Chu, as presented in [71], did not work. When this was pointed out, Chu provided an adjustment in which the parameter μ in [71, (2.2)] was replaced by $\mu\lambda_1/|\lambda_1|$. With this adjustment, it was possible to solve 4 by 4 and 5 by 5 test cases that previously caused failure. The results reported in this section use the adjusted algorithm.

Both Matlab routines GTD (see Appendix) and SVD_EIG [71] require $O(n^2)$ flops, so in an asymptotic sense, the approaches are equivalent. In Table 6–1 we compare the actual running times of GTD and SVD_EIG for matrices of various dimensions. These computer runs were performed on a Sun Workstation with 1 GB memory. In making these runs, the portion of the GTD code connected with the updating of the matrices $\mathbf{P}$ and $\mathbf{Q}$ was deleted since SVD_EIG does not accumulate the unitary matrices. The input arrays $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ were generated in the following way: Using the Matlab routine RAND, we randomly generated a square matrix whose element lie between 0 and 1. The

singular values $\boldsymbol{\sigma}$ were computed using the Matlab routine SVD and the eigenvalues $\boldsymbol{\lambda}$ were computed using Matlab's EIG. By the theorem of Weyl [69], $\boldsymbol{\lambda} \preceq \boldsymbol{\sigma}$. We then used both SVD_EIG and GTD to generate matrices with the specified singular values and eigenvalues. Five different matrices of each dimension were generated and the average running time is reported in Table 6–1.

The times shown in Table 6–1 indicate that GTD becomes increasingly more efficient than SVD_EIG as the matrix dimension increases. For a dimension of 100, GTD is about three times faster than SVD_EIG. For a dimension of 1600, GTD is about 70 times faster than SVD_EIG. In an efficiently designed compiled code, the difference in speed between these two approaches to inverse eigenvalue problems could be more substantial: the permutations appearing in GTD could be replaced by the updating of a pointer array; also, the columns of $\mathbf{R}$ that are zero except for the diagonal entry could be flagged, and when multiplying $\mathbf{R}$ by $\mathbf{G}_1$, we can skip the multiplication of these essentially zero columns.

In Table 6–1 we also compare the specified singular values and eigenvalues to those obtained by applying Matlab's SVD and EIG routines to the generated matrices. That is, for each matrix output by either SVD_EIG or GTD, we use Matlab's routines to compute the singular values and eigenvalues, and we compare to the specified singular values and eigenvalues using the sup-norm. The errors reported in Table 6–1 are the average errors for the 5 random matrices of each dimension. Both routines generate matrices with singular values that match those computed by Matlab's SVD routine to within 13 or 14 digits. Observe that GTD always matches exactly the prescribed eigenvalues since the generated matrix is triangular, with the specified eigenvalues on the diagonal. The error in the eigenvalues of the matrix generated by SVD_EIG was comparable to the singular value error for matrices of dimension up to 400. Thereafter, the error in the eigenvalues grew quickly. When the matrix dimension doubled from 400 to 800, the error increased roughly by the factor 10^3. And when the matrix dimension doubled again from 800 to 1600, the error increased roughly by the factor 10^6.

A recursive algorithm can require a significant amount of memory. While SVD_EIG executed, we monitored the memory usage using the Unix "top" command. We observed that for a matrix of dimension 1600, the memory consumption grew to 319 MB. Since a complex double precision matrix of dimension 1600 occupies about 41 MB memory, the recursion required more than 7 times as much space as the matrix itself.

6.4 Conclusions

In this chapter, we introduce two novel matrix decomposition algorithms, including the geometric mean decomposition and the generalized triangular decomposition. The GMD $\mathbf{H} = \mathbf{QRP}^*$ is a solution of the maximin problem (6.3); the smallest diagonal element of $\mathbf{R}$ is as large as possible. Starting with the SVD, we show that the GMD can be computed using a series of Givens rotations, and row and column exchanges. Alternatively, the GMD could be computed directly, without performing an initial SVD, if $\mathbf{H}$ is first reduced by unitary transformations to a real matrix. In a further extension of our algorithm for the GMD, we show in [51] how to compute a factorization $\mathbf{H} = \mathbf{QRP}^*$ where the diagonal of $\mathbf{R}$ is any vector satisfying the Weyl multiplicative majorization conditions [69]. The GTD represents the most general unitary decomposition $\mathbf{H} = \mathbf{QRP}^*$. That is, the diagonal $\mathbf{r}$ of $\mathbf{R}$ must satisfy $\mathbf{r} \preceq \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the vector of singular values for $\mathbf{H}$, while for any diagonal $\mathbf{r}$ with $\mathbf{r} \preceq \boldsymbol{\sigma}$, we can write $\mathbf{H} = \mathbf{QRP}^*$. The GTD includes, as special cases, the singular value decomposition, the Schur decomposition, the $\mathbf{QR}$ decomposition, and the GMD. Similar to GMD, given the SVD, the GTD based on $\mathbf{r}$ can also be evaluated using a series of Givens rotations and permutations. The GTD algorithm provides a new proof of Horn's theorem [70]. Applications of the GTD include transceiver design for MIMO communications [72] and inverse eigenvalue problems surveyed extensively in [73].

Appendix
Matlab Implementation of GTD

```
% Input:
%
% H = U*S*V' (singular value decomposition of H)
%         U and V orthonormal columns
%         S diagonal matrix with nonnegative diagonal entries
%         r desired diagonal of R
%          --nnz (r) = nnz (S)
%          --r multiplicatively majorized by diag (S)
%          --product nonzero r = product nonzero diag (S)
%
% Output:
%
```

```
% H = Q*R*P' (GTD based on r)
%         P and Q orthonormal columns
%         R upper triangular, R (i, i) = r (i)

function [Q, R, P] = gtd (U, S, V, r)

d = diag (S) ;
K = min (size (S)) ;
P = V ; Q = U ; R = zeros (K) ;
for k = 1 :  K-1
     rk = r (k) ;
     abs_rk = abs (rk) ;
     kp1 = k + 1 ; km1 = k - 1 ;
     I = find (abs (d (k :  K)) > abs_rk) ;
     if ( isempty (I) )
          [x, p] = max (abs (d (k :  K))) ;
          p = p + km1 ;
     else
          I = I + km1 ;
          [x, p] = min (abs (d (I))) ;
          p = I (p) ;
     end
     delta1 = d (p) ;
     d ([k p]) = d ([p k]) ;
     I = find (abs (d (kp1 :  K)) <= abs_rk) ;
     if ( isempty (I) )
          [x, q] = min (abs (d (kp1 :  K))) ;
          q = q + k ;
     else
          I = I + k ;
          [x, q] = max (abs (d (I))) ;
          q = I (q) ;
     end
```

```
delta2 = d (q) ;
d ([kp1 q]) = d ([q kp1]) ;
sq_delta1 = abs (delta1)^2 ;
sq_delta2 = abs (delta2)^2 ;
sq_rk = abs_rk^2 ;
denom = sq_delta1 - sq_delta2 ;
if ( (denom <= 0) | (sq_rk > sq_delta1) )
     c = 1 ; s = 0 ;
elseif ( sq_rk < sq_delta2 )
     c = 0 ; s = 1 ;
else
     c = sqrt ((sq_rk - sq_delta2)/denom) ;
     s = sqrt (1-c*c) ;
end
if ( sq_rk > 0 )
     x = -s*c*rk*denom/sq_rk ;
     y = delta1*delta2*rk/sq_rk ;
     G1 = [ c -s
            s  c ] ;
     G2 = [ c*delta1 -s*(delta2')
            s*delta2  c*(delta1') ] ;
     G2 = ((rk')/sq_rk) * G2 ;
else
     x = 0.  ;
     y = delta1 ;
     G1 = [ 0 -1
            1  0 ] ;
     G2 = G1 ;
end
if ( k > 1 )
     % permute the columns
     R (1:km1, [k p]) = R (1:km1, [p k]) ;
     R (1:km1, [kp1 q]) = R (1:km1, [q kp1]) ;
```

```
        % apply G1 to R
        R (1:km1, [k kp1]) = R (1:km1, [k kp1])*G1 ;
    end
    R (k, k) = rk ;
    R (k, kp1) = x ;
    d (kp1) = y ;

    % permute the columns
    P (:, [k p]) = P (:, [p k]) ;
    P (:, [kp1 q]) = P (:, [q kp1]) ;
    Q (:, [k p]) = Q (:, [p k]) ;
    Q (:, [kp1 q]) = Q (:, [q kp1]) ;

    % apply G1 to P
    P (:, [k kp1]) = P (:, [k kp1])*G1 ;

    % apply G2 to Q
    Q (:, [k kp1]) = Q (:, [k kp1])*G2 ;
end

R (K, K) = r (K) ;
if ( r (K) ~= 0 )
    Q (:, K) = Q (:, K)*d (K)/ r (K) ;
end
P = P (:, 1:K) ;
Q = Q (:, 1:K) ;
```

CHAPTER 7
CONCLUSIONS

This dissertation studies the signal processing aspect of MIMO communications. We present a new perspective to the MIMO communications: any MIMO scheme can be regarded as a MIMO channel decomposer, which decomposes a MIMO channel into multiple scalar subchannels. Based on this perspective, this dissertation presents three novel MIMO transceiver designs: the geometric mean decomposition (GMD) scheme, the uniform channel decomposition (UCD) scheme, and the tunable channel decomposition (TCD) scheme. All these schemes deploying either a VBLAST detector at the receiver or a dirty paper precoder at the transmitter. These transceiver designs represent a paradigm shift from the conventional linear MIMO transceiver design to the nonlinear ones. The superior performance of the GMD and UCD scheme unveils the practical significance of the collaborations between the transmitter and receiver. That is, such collaborations facilitate achieving the optimal tradeoff between the diversity and multiplexing gains promised by the MIMO communication theory. The TCD scheme represents a unifying solution to a considerably wide range of problems, including designing the precoder for OFDM communications and the optimal CDMA sequences.

Motivated by the application of transceiver designs, this dissertation also introduces two novel matrix decomposition algorithms, i.e., the geometric mean decomposition (GMD) and the generalized triangular decomposition (GTD). The two matrix decompositions are the cornerstones of the three transceiver designs proposed in this dissertation. Moreover, the two matrix decomposition algorithms have significant implications in the matrix analysis community. For instance, the GTD is a new and more efficient solution to the inverse eigenvalue problem.

REFERENCES

[1] I. E. Telatar, "Capacity of multiple antenna Gaussian channels," *AT&T Technical Memorandum*, June 1995.

[2] G. J. Foschini, Jr., "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. Journal*, vol. 1, pp. 41–59, Autumn 1996.

[3] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[4] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1841–1852, November 1999.

[5] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multiple-element arrays," *Wireless Personal Communications*, vol. 6, pp. 311–335, March 1999.

[6] M. Sellathurai and S. Haykin, "Turbo-BLAST for wireless communications: Theory and experiments," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2538–2545, October 2002.

[7] M. Sellathurai and G. J. Foschini, "Stratified diagonal layered space-time architectures: signal processing and information theoretic aspects," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2943–2954, November 2003.

[8] G. J. Foschini, D. Chizhik, M. J. Gans, C. Papadias, and R. A. Valenzuela, "Analysis and performance of some basic space time architectures," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 303–320, April 2003.

[9] B. Hassibi, "A fast square-root implementation for BLAST," *Thirty-Fourth Asilomar Conf. Signals, Systems and Computers*, pp. 1255–1259, November 2000.

[10] S. L. Ariyavisitakul, "Turbo space-time processing to improve wireless channel capacity," *IEEE Transactions on Communications*, vol. 48, pp. 1347–1358, August 2000.

[11] J. Benesty, Y. Huang, and J. Chen, "A fast recursive algorithm for optimum sequential signal detection in a BLAST system," *IEEE Transactions on Signal Processing*, vol. 51, pp. 1722–1730, July 2003.

[12] S. M. Alamouti, "A simple transmit diversity techniques for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1451–1458, October 1998.

[13] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communications: Performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, pp. 744–765, March 1998.

[14] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, pp. 1456–1467, July 1999.

[15] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Combined array processing and space-time coding," *IEEE Transactions on Information Theory*, vol. 47, pp. 199–207, Feb. 1999.

[16] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Transactions on Information Theory*, vol. 49, pp. 1073–1096, May 2003.

[17] *Available Online: http://www.3gpp.org.*

[18] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems," *IEEE Transactions on Communications*, vol. 42, pp. 3221–3231, December 1994.

[19] G. G. Raleigh and J. M. Cioffi, "Spatial-temporal coding for wireless communication," *IEEE Transactions on Communications*, vol. 46, pp. 357–366, March 1998.

[20] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Filterbank transceiver optimizating information rate in block transmissions over dispersive channels," *IEEE Transactions on Information Theory*, vol. 45, pp. 1019–1032, April 1999.

[21] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers part i: Unification and optimal designs," *IEEE Transactions on Signal Processing*, vol. 47, pp. 1988–2006, July 1999.

[22] H. Sampath, P. Stoica, and A. Paulraj, "Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion," *IEEE Transactions on Communications*, vol. 49, pp. 2198–2206, December 2001.

[23] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1051–1064, May 2002.

[24] E. Onggosanusi, A. Sayeed, and B. V. Veen, "Optimal antenna diversity signaling for wideband space-time wireless channels utilizing channel state information," *IEEE Transactions on Communications*, vol. 50, pp. 341–353, February 2002.

[25] E. Onggosanusi, A. Sayeed, and B. V. Veen, "Efficient signaling schemes for wideband space-time wireless channels using channel state information," *IEEE Transactions on Vehicular Technology*, vol. 52, pp. 1–13, January 2003.

[26] D. Palomar, J. Cioffi, and M. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2381–2401, September 2003.

[27] D. Palomar and M. Lagunas, "Joint transmit-receive space-time equalization in spatially correlated MIMO channels: A beamforming approach," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 730–743, June 2003.

[28] D. Palomar, M. Lagunas, and J. Cioffi, "Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1179–1197, May 2004.

[29] L. Collin, O. Berder, P. Rostaing, and G. Burel, "Optimal minimum distance-based precoder for MIMO spatial multiplexing systems," *IEEE Transactions on Signal Processing*, vol. 52, pp. 617–627, March 2004.

[30] J.-K. Zhang, A. Kavčić, X. Ma, and K. M. Wong, "Design of unitary precoders for ISI channels," in *Proceedings IEEE International Conference on Acoustics Speech and Signal Processing*, vol. III, (Orlando, Florida), pp. 2265–2268, 2002.

[31] M. Rupf and J. L. Massey, "Optimal sequence multisets for synchronous code-division multiple-access channels," *IEEE Transactions on Information Theory*, vol. 40, pp. 1261–1266, July 1994.

[32] D. W. Bliss, K. W. Forsythe, A. O. Hero, and A. F. Yegulalp, "Environmental issues for MIMO capacity," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2128–2142, September 2002.

[33] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1997.

[34] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge: Cambridge University Press, 1985.

[35] S. Haykin, *Adaptive Filter Theory.* Englewood Cliffs, NJ: Prentice-Hall, Third Edition, 1996.

[36] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York, NY: John Wiley and Sons, Inc., 1968.

[37] G. Ginis and J. M. Cioffi, "On the relationship between V-BLAST and the GDFE," *IEEE Communications Letters*, vol. 5, pp. 364–366, September 2001.

[38] G. Ginis and J. M. Cioffi, "Vectored transmission for digital subscriber line systems," *IEEE Journal on Selected Areas in Communications*, vol. 20, pp. 1085–1104, June 2002.

[39] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, pp. 1691–1706, July 2003.

[40] G. Ginis and J. M. Cioffi, "A multi-user precoding scheme achieving crosstalk cancellation with application to DSL systems," *Proc. 34th Asilomar Conference Signals, Systems, Computers,* Asilomar, CA, vol. 2, pp. 1627–1631, 29 Oct.-1 Nov. 2000.

[41] M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983.

[42] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electron. Lett.*, vol. 7, pp. 138–139, March 1971.

[43] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Communications*, pp. 774–780, August 1972.

[44] W. Yu and J. M. Cioffi, "Trellis precoding for the broadcast channel," *Global Telecommunications Conference*, vol. 2, pp. 1344–1348, November 2001.

[45] B. C. Banister and J. Zeidler, "Feedback assisted transmission subspace tracking for MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 452–463, April 2003.

[46] A. Poon, D. Tse, and R. W. Brodersen, "An adaptive multiantenna transceiver for slowly flat fading channels," *IEEE Transactions on Communications*, vol. 51, pp. 1820–1827, November 2003.

[47] B. Hassibi, "An efficient square-root implementation for BLAST," *http://cm.bell-labs.com/who/hochwald/papers/squareroot/*, 2000.

[48] M. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1405 – 1409, Nov 2-5 1997.

[49] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, pp. 1912–1921, August 2003.

[50] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Available: http://inst.eecs.berkeley.edu/∼ee224b/sp04/#Course Notes, 2004.

[51] Y. Jiang, W. Hager, and J. Li, "The generalized triangular decomposition," *SIAM Journal on Matrix Analysis and Applications*, Available online: http://www.sal.ufl.edu/yjiang/papers/gtd.pdf Submitted.

[52] Y. Jiang and J. Li, "Adaptable channel decomposition for MIMO communications," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, March 18 – 23, 2005.

[53] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. second edition, John Wiley and Sons, Inc., 1984.

[54] P. Viswanath and V. Anantharam, "Optimal sequences and sum capacity of synchronous CDMA systems," *IEEE Transactions on Information Theory*, vol. 45, pp. 1984–1991, September 1999.

[55] P. Viswanath, V. Anantharam, and D. Tse, "Optimal sequences, power control and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers," *IEEE Transactions on Information Theory*, vol. 45, pp. 1968–1983, September 1999.

[56] T. Guess, "Optimal sequence for CDMA with decision-feedback receivers," *IEEE Transactions on Information Theory*, vol. 49, pp. 886–900, April 2003.

[57] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc, 1991.

[58] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization*. New York: Academic, 1979.

[59] Y. Jiang, J. Li, and W. Hager, "Uniform channel decomposition for MIMO communications," *IEEE Transactions on Signal Processing*, available online: http://www.sal.ufl.edu/yjiang/papers/ucdR3.pdf, to appear 2004. The conference version presented at Asilomar conference, Nov. 2004.

[60] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge: Cambridge University Press, 1991.

[61] E. Beltrami, "Sulle funzioni bilineari," *Giornale De Matematiche*, vol. 11, pp. 98–106, 1873.

[62] C. Jordan, "Mémoire sur les formes bilinéaires," *J. Math. Pures Appl.*, vol. 19, pp. 35–54, 1874.

[63] I. Schur, "On the characteristic roots of a linear substitution with an application to the theory of integral equations," *Math. Ann.*, vol. 66, pp. 488–510, 1909.

[64] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*. Philadelphia: SIAM, 1996.

[65] P. Kosowski and A. Smoktunowicz, "On constructing unit triangular matrices with prescribed singular values," *Computing*, vol. 64, pp. 279–285, 2000.

[66] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solution," *Numerische Mathematik*, vol. 14, pp. 403–420, 1970.

[67] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, 1983.

[68] J. H. Wilkinson and C. Reinsch, "Linear algebra," in *Handbook for Automatic Computation* (F. L. Bauer, ed.), vol. 2, (Berlin), Springer-Verlag, 1971.

[69] H. Weyl, "Inequalities between two kinds of eigenvalues of a linear transformation," *Proc. Nat. Acad. Sci. U. S. A.*, vol. 35, pp. 408–411, 1949.

[70] A. Horn, "On the eigenvalues of a matrix with prescribed singular values," *Proc. Amer. Math. Soc.*, vol. 5, pp. 4–7, 1954.

[71] M. T. Chu, "A fast recursive algorithm for constructing matrices with prescribed eigenvalues and singular values," *SIAM J. Numer. Anal.*, vol. 37, pp. 1004–1020, 2000.

[72] Y. Jiang, J. Li, and W. Hager, "Joint transceiver design for MIMO communications using geometric mean decomposition," *IEEE Transactions on Signal Processing*, to appear. Available online: http://www.sal.ufl.edu/yjiang/papers/gmdCommR2.pdf.

[73] M. T. Chu, "Inverse eigenvalue problems," *SIAM Rev.*, vol. 40, no. 1, pp. 1–39 (electronic), 1998.

BIOGRAPHICAL SKETCH

Yi Jiang was born in Yixing, Jiangsu Province, China, in November 1978. He received his B.Sc. degree in Electronic Engineering and Information Science from the University of Science and Technology of China (USTC) in July 2001, the M.S. degree in Electrical Engineering from the University of Florida (UF) in May 2003.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jian Li, Chair
Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Kenneth K. O
Professor of Electrical and Computer
Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

John M. Shea
Assistant Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Tan F. Wong
Associate Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

William W. Hager
Professor of Mathematics

TRANSCEIVER DESIGN FOR MIMO COMMUNICATIONS
– A CHANNEL DECOMPOSITION PERSPECTIVE

Yi Jiang
(352) 392-5241
Department of Electrical and Computer Engineering
Chair: Jian Li
Degree: Doctor of Philosophy
Graduation Date: May 2005

It is recently discovered that deploying multiple transmitting and multiple receiving antennas in a wireless communication system can drastically improve the data rate and reliability of wireless communications, even without consuming additional bandwidth and input power. This so-called multi-input multi-output (MIMO) technology has been under intense research and will be applied to the next generation of wireless communication networks.

This dissertation focuses on designing practical transceiver designs for MIMO systems with sound theoretical foundations. Three designs, i.e., the GMD, UCD, and TCD schemes, are proposed. These designs represent a paradigm shift from the conventional linear designs to the nonlinear designs while keeping the implementation complexity low. It is proven, both through theoretical analyses and numerical simulations, that the three designs are much better than their linear counterparts in that they can achieve faster and more reliable communications. The schemes proposed in this dissertation will probably play an important role in the next generation wireless fidelity (Wi-Fi) and digital subscribe line (DSL) technologies. Besides its engineering significance, this dissertation also invents two matrix decompositions, which are significant contributions to the numerical analysis community.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jian Li, Chair
Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Kenneth K. O
Professor of Electrical and Computer
Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

John M. Shea
Assistant Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Tan F. Wong
Associate Professor of Electrical and
Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

William W. Hager
Professor of Mathematics

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 2005

Pramod P. Khargonekar
Dean, College of Engineering

Kenneth Gerhardt
Dean, Graduate School